Industrial Statistics

Ronald Christensen and Aparna V. Huzurbazar
Department of Mathematics and Statistics
University of New Mexico

To Walt and Ed

Contents

Preface

Industrial Statistics is largely devoted to achieving, maintaining, and improving quality. Industrial Statistics provides tools to help in this activity. But *management is in charge of the means of production, so only management can achieve, maintain, or improve quality.*

In the early 20th century Japan was renown for producing low quality products. After World War II, with influence from people like W. Edwards Deming and Joseph M. Juran, the Japanese began an emphasis on producing high quality goods and by the 1970s began out-competing American automobile and electronics manufacturers. Deming had had previous unsuccessful experiences in America with implementing (statistical) quality programs and had decided that the key was getting top management to buy into an overall program for quality.

While *the basic ideas of quality management are quite stable*, actual implementations of those ideas seem subject to fads. When I first become interested in Industrial Statistics, Total Quality Management (TQM) was all the rage. That was followed by Six-Sigma which seems to have run its course. Lean seems to have been next up and even that seems to be passing. Lean Six-Sigma is what I see being pushed most in 2021. I expect a new fad to arise soon.

Wikipedia has separate entries for Quality Management, Quality Management Systems (QMS), and TQM. I cannot tell that there is much difference between them other than what terminology is in vogue. The American Society for Quality (ASQ) has four major topics on their list for learning

- Quality Management
- Standards
- Six-Sigma
- Lean

0.1 Standards

Standards refers to standards for quality management. The International Organization for Standardization (ISO - *not* an acronym in any of English, French or Russian) produces the 9000 series standards for Quality Management. The key document seems to be ISO 9001:2015 which specifies requirements for quality management. This was reviewed and confirmed in 2021. Other key standards are ISO 9000:2015 on fundamentals and vocabulary and ISO 9004:2018 for continuous improvement of quality. See https://www.iso.org/iso-9001-quality-management.html and https://www.iso.org/standard/62085.html.

0.2 Six-Sigma

Six-Sigma is an entire management system that began at Motorola and was famously exploited by General Electric (GE). It is named after a very good idea related to control charts but has been extrapolated far beyond that.

The basic idea of control charting is that with a process that is under control, virtually all of the output should fall within three standard deviations of the mean. (For technical reasons when determining control status this idea is best applied to the average of small groups of observations that are combined for some rational reason [*rational subgroups*].) Indeed, this idea is the basis for an *operational definition* of what it means to have a process under control. (For those with

some probability background you can think of it as an operational definition of what it means to be *independent and identically distributed (iid)*.) The interval from the mean minus three standard deviations up to the mean plus three standard deviations is known as the *process capability* interval. Typically a product has some *specification interval* within which the product is required to fall. If the process capability interval is contained within the specification interval we are good to go. The standard deviation is typically denoted by the Greek letter sigma (σ).

The fundament idea behind Six-Sigma is striving to get the much larger interval from the mean minus six standard deviations (six-sigma) up to the mean plus six standard deviations within the specification interval. To do this may require cutting product variability in half. In such a case, if your process is on target (the middle of the specification interval) you will have very little variability and even if your process strays off of the target a bit, you can remain within the specification limits. But the overall Six-Sigma program is vastly more complicated.

David Wayne (http://q-skills.com/Deming6sigma.htm) says "Six Sigma, while purporting to be a management philosophy, really seems more closely related to Dr. Joseph Juran's more project-oriented approach, with a deliberate, rigorous technique for reaching a problem resolution or an improvement. Dr. Deming's approach is more strategic, theoretical and philosophical in nature, and does not carry the detailed explicitness of the Six Sigma approach." My memory (I am still looking for an exact reference) is that Deming was critical of Six-Sigma for being overly focused on financial issues. (Not surprisingly, this emphasis on financial issues seems to have made Six-Sigma more popular with top management.)

Hahn, Hill, Hoerl, and Zinkgraf (1999) and Montgomery and Woodall (2008) present overviews of Six-Sigma from a statistical viewpoint. The panel discussion Stenberg et. al. (2008) also discusses Six-Sigma quite a bit.

0.3 Lean

Lean is a program for eliminating waste based on Toyota's program for doing so. It is often presented as a cycle similar to the Shewhart Cycle: Plan, Do, Study, Act that is discussed in Section 1.3. The lean version is: Identify Value, Map the Value Stream, Create Flow, Establish Pull, Seek Perfection. The key ideas are to minimize steps in your process that do not add value and to implement steps that increase value. For more information see https://www.lean.org/WhatsLean/Principles.cfm or cips-lean

0.4 This Book

Industrial Statistics obviously can involve any and all standard statistical methods but it places special emphasis on two things, control charts and experimental design. Here we review a wide range of statistical methods, discuss basic control charts, and introduce industrial experimental design. The seminal work in modern Industrial Statistics is undoubtedly Walter Shewhart's 1931 book *Economic Control of Quality of Manufactured Product*. In my decidedly unhumble opinion, the two best Statistics books that I have read both relate to industrial statistics. They are Shewhart's (1939) *Statistical Method from the Viewpoint of Quality Control* (heavily edited by W. Edwards Deming) and D. R. Cox's (1958) *Planning of Experiments*. Within experimental design Industrial Statistics places special emphasis on screening designs and on response surface methodologies. My *Topics in Experimental Design (TiD)* (https://www.stat.unm.edu/~fletcher/TopicsInDesign) discusses these subfields but is written at a higher mathematical level.

When I joined ASQ it seemed to be a professional organization similar to the American Statistical Association. Now it seems to me that they are primarily in the business of certifying quality professionals and selling materials to facilitate certification. Relative to their test for Manager of Quality/Organizational Excellence Certification CMQ/OE, this book covers much of Section IV (Quality Management Tools) and a bit of Section IIIe (Quality Models and Theories). Other ASQ certifications whose bodies of knowledge have some crossover are Certified Quality Engi-

neer (CQE), Certified Six Sigma Black Belt (CSSBB), Certified Six Sigma Green Belt (CSSGB), Certified Reliability Engineer (CRE), and Certified Quality Inspector (CQI).

0.4.1 Computing

The computing is done in Minitab because it is the simplest package I know. Minitab began as a general statistical package but long ago oriented itself towards industrial applications. My introduction to Minitab is available at `www.stat.unm.edu/~fletcher/MinitabCode.pdf`. It was written as a companion to Christensen (2015). Chapters 1 and 3 are the most important. (It also contains an intro to SAS.) Last I looked, you could get a six month Minitab academic license for $33 at `estore.onthehub.com`. They also had a one-year $55 license. (This book does *not* use the *Minitab Workspace* package!) I have no personal experience with it but JMP (a SAS product) seems comparable to Minitab in ease of use and industrial orientation.

You can do pretty much anything in statistics using the free programming language R. There are a number of R packages that address control charts among these are `qcc`, qcr (which uses `qcc`), `qicharts`, `qicharts2` , `ggQC` (quality control charts for ggplot). My introduction to R is available at `www.stat.unm.edu/~fletcher/Rcode.pdf`. It has the same structure as my Minitab introduction. I will discuss R only a little.

The data used here can be accessed from `www.stat.unm.edu/~fletcher/industrial-data.zip`. The data in Table x.y is in file `tabx-y.dat`. FYI: references to qexx-yyy that occasionally occur are to data found in *Quality Engineering*, 19xx, page yyy.

Ronald Christensen
Albuquerque, New Mexico
July, 2021

Chapter 1

Introduction

Most of this book deals with statistical tools used to establish and improve the quality of industrial (and service sector) processes. These are good and useful tools, but they are only tools. Statistical procedures are helpful and, when considered in the broadest sense, perhaps even indispensable to achieving, maintaining, and improving quality. Without an appreciation for data and the variability inherent in data, only the simplest of quality problems can be solved. Nonetheless, statistical methods are only tools, they are not panaceas. Ultimately, management is responsible for quality. Management owns the industrial processes and only management can improve them. Only a management that is committed to creating high quality products will achieve this goal. And even then, it will only be the management teams that also have the knowledge base to execute their goals that will be successful. In this chapter we briefly present some ideas on the kind of management necessary to achieve high quality on a continuing basis. The remainder of the book focuses on statistical tools.

EXERCISE 1.0.1 Read the article https://williamghunter.net/george-box-articles/the-scientific-context-of-quality-improvement to get the viewpoint of two famous statisticians on quality improvement. It briefly surveys many of the topics we will discuss. (There is a video presentation of this material that is linked to a video in a (much) later assignment.)

1.1 Four Principles for Quality

Around the middle of the twentieth century, W. Edwards Deming played an instrumental role in convincing Japanese businesses to emphasize quality in production. In the last half of the twentieth century, Japan became an industrial giant. Statistics plays a vital role in the production of high quality goods and services. In discussing quality in the production of goods and services, one must always remember that quality does not exist without consideration of price. Before getting an Apple Watch and smart phone, for many years I was very content with my Pulsar wristwatch. It kept time better than I needed and the cost was very reasonable. On the other hand, a Rolex looks more spectacular and may have kept better time. Unfortunately, with my salary, the improved quality of a Rolex was not sufficient to offset the increased cost. Similarly, I doubt that I will ever be in the market for one of those wonderful cars made by Rolls-Royce. Nonetheless, consumers care deeply about the quality of the goods and services that they are able to purchase. Ishikawa (1985, p. 44) defines the goal of quality production, "To practice quality control is to develop, design, produce and service a quality product which is most economical, most useful, and always satisfactory to the consumer." Goods and services must be produced with appropriate quality at appropriate prices and in appropriate quantities.

This chapter examines some of Deming's ideas about business management and the production of high quality goods and services. Deming (1986) indicates that the proper goal of management is to stay in business. Doing so provides jobs, including management's own. He argues that, to a large extent, profits take care of themselves in well run businesses. Well run businesses are those that produce high quality goods and services and that *constantly improve* the quality of their goods and services.

Deming's (1986) quality management program involves 14 Points (P1 – P14), Seven Deadly Diseases (DD1 – DD7), plus obstacles to quality improvement. We cannot remember that much so we have reduced his ideas to four primary principles:

1. Institute and Maintain Leadership for Quality Improvement,
2. Create Cooperation,
3. Train, Retrain, and Educate,
4. Insist on Action.

In the next subsections, these are each discussed in turn. Deming (1993) is an easy to read introduction to Deming's ideas. Deming (1986) is more expansive and detailed. Walton (1986) provides a nice introduction to Deming's ideas. (I've been trying unsuccessfully to get my wife to read Walton for over a decade.)

1.1.1 Institute and Maintain Leadership for Quality Improvement

There is a process, instituted and maintained by management, for creating and improving new and existing products and services. It is this process that determines quality and ultimately business success. Developing a high quality product and service is not enough. *The Process of Production and Service Must be Continually Improved* to keep ahead of the competition (Deming's P5). Note our use of the pair "product and service." Service is the obvious product in many companies but *serving the needs of the customer is really the product in all companies*. Everybody has a customer. The first job in quality improvement is to identify yours. (Are students customers or products? Prior to the internet, college textbooks were clearly written for instructors and not the more obvious target, students.)

Quality improvement is not easy; you have to persist in your efforts. There are no quick fixes. (Deming says there is no "instant pudding.") Deming's P2 is to *Adopt the New Philosophy*. Patching up old methods is not sufficient to the task. Maintaining *Constancy of Purpose* is Deming's P1. Lack of constancy of purpose is also his DD1. Constancy of purpose is impossible with *Mobility of Top Management*, Deming's DD4. Managers who are not there for the long haul cannot focus on long term objectives, like quality. Mobile managers need to look good in the short term so they can move on to wrecking the next business. The lack of constancy of purpose leads to a debilitating *Emphasis on Short Term Profits*, Deming's DD2.

For a birthday present we bought a gift certificate at an Albuquerque T-shirt shop. Twenty dollars to buy a twenty dollar certificate. When the recipient went to the store, he brought a coupon for 25% off the price of a shirt. The coupon, as is often the case, was not good with other special offers. The manager of the store chose to view the use of a gift certificate as a special offer. The manager received full price on the T-shirt bought that day; they maximized their short term profit. But we never bought anything else from them and we discouraged our friends from patronizing them. Apparently other people had similar experiences; the company is now out of business.

It is not enough to satisfy the customers current needs. It is not enough to produce current goods and services of high quality. Management must lead the customer. Management must think about what the company should be doing five years from now. They must anticipate what their customers will need five years from now. Producers should have a better idea of where their industry is going than their customers. If a competitor provides the next improvement, your customer will have to switch to maintain her competitive position. This doesn't mean rushing into the market with a low quality product. That is not the way to long term success. It means entering the market in a timely fashion with a high quality product. Improvement requires innovation. Innovation requires research. Research requires education. Often, innovation comes from small focused groups rather than large amorphous research institutions.

Improved quality does not just happen. It requires a program to succeed and a program requires leadership. If your quality could improve without a program, why hasn't it improved already? Dem-

ing's P7 is to *Institute Leadership*. Leadership requires a wide view of the business environment. Just because you cannot measure something does not mean it is unimportant. It is easy to measure the increased costs of a quality program but it is impossible to measure the increased revenue derived therefrom. In general, many financial features are easy to measure. They are not all important. Financial measures lead to shipping product regardless of quality. Quality, on the other hand, is hard to measure. *Running a company on visible figures alone* is Deming's DD5.

High quality means dependability; quality improvement means reducing variability in the process of producing goods and services. First you need to establish that there is a market for your product and service. Then you need to focus on doing well what you have chosen to do.

Until the mid 1970s Arby's made a great roast beef sandwich, *some of the time*. Unfortunately, getting an almost unchewable sandwich was not a rare event. It is common knowledge that the key to success in the fast food business is uniformity of product. You serve the same basic food every time to every customer from Fairbanks to Key West. Arby's had a problem. They solved their variability problem by switching to roasts made of pressed beef. Note that they did not find a better way to serve the same product; they switched to a new product that had less variability. The chances of getting a sandwich with tough meat are now very small but reducing the variability would have served no purpose if nobody wanted to eat pressed beef roasts. As Arby's is still in business, they must have a market. It is not the same market they had before because it is not the same product they had before. (I, for one, used to be a regular customer but have hardly set foot in an Arby's for 45 years.) But Arby's is undoubtedly serving their new market better than they served their old one. Their customers know what they are going to get and go away contented with it.

Before one can improve the production of goods and services, the current process of production must reach a stable point. If you perform the job one way this month and a different way next month, then you don't have any process of production. You must have a standard way of doing things before you can begin to improve the way you do things. Statistical control charts are used 1) to establish whether a system of production exists, 2) to display the variability built into the process, and 3) to identify *special causes* that have disrupted the system.

Quality needs to be designed into the product and the process of making the product. A rule of thumb (see Deming, 1986, p. 315) is that 94% of all problems are due to the system of production, something only management can alter. Only 6% of problems are due to special causes that workers may be able to control. It is obvious that if you have trouble with everybody's work, i.e., if nobody can accomplish the job the way you want it done, the problem must be in what you are asking people to do. Management needs to find processes that allow everybody to accomplish the job successfully. If you are unhappy with your workers and think you can solve your problem by getting new ones, you are just kidding yourself. The pool of workers is stable; you need to improve your systems. *Leadership is taking the initiative to help people do their jobs better. Workers believe in quality. Managers are the ones who sacrifice quality for short term profits*.

1.1.2 Create Cooperation

Nearly everyone agrees that people are an organization's greatest asset but few use that asset really effectively. To use people effectively, management must *Drive Out Fear* (P8) and in other ways *Remove Barriers to Pride of Workmanship* (P12). Innovation and improvement require communication; communication must be actively encouraged. Management must *Break Down Barriers to Communication* (essentially P9).

Human life is a paradox of cooperation and competition. In questions of survival, people do both. They are forced to compete with those that threaten them. They cooperate with other people who are threatened by a common danger. Responses to competition frequently become dysfunctional if the competition is too desperate. The trick is to foster cooperation within the organization and to focus competition externally. If you are competing with another employee to survive within the organization, you cannot cooperate with that employee for the good of the organization. Your own needs will come first.

When their survival is not threatened, people still compete with each other but on a tamer level and not to the exclusion of productive, cooperative achievement. We need people competing to be the most valuable player on a team rather than a hot-shot self-centered superstar. We need Larry Birds and Magic Johnsons: people who's greatness stemmed from making their teammates better. (Maybe LeBron Jameses?)

Driving Out Fear (P8) starts with providing job security. If a person cannot perform adequately in one job, find them another. The fear of failure is a huge barrier. Failure and mistakes are necessary for innovation. If you can get an unpleasant job assignment or lose your raise, promotion, or job for trying something new or making "annoying" suggestions, you won't do it. In a climate of fear, you cannot even find out what is going on in the organization because people fudge the figures out of fear. If top management threatens to fire everyone in a shop if the shop ever exceeds 10% defectives, you can be sure that nobody in the shop will ever tell management that defectives have exceeded 10% and management will never know the true percentage of defectives. You can buy a person's time but you have to earn their loyalty and confidence. After years of managing by fear, driving fear out can be a long process.

Management must *Remove Barriers to Pride of Workmanship* (P12). This begins with simple measures such as ensuring that tools and machines work properly. It begins by allowing people to do their jobs correctly. But management must also remove the barriers that they have intentionally set up. Setting goals without a program to meet them does not help anyone. To *Eliminate Numerical Quotas* is Deming's P11. If all your effort is devoted to producing 100 units per day, you have no effort left for ensuring quality. Raising an already high quota is a guarantee of low quality. The least damaging quotas are those that everyone can meet. Quotas also stifle effort and encourage standing around because "I met my quota for the day."

Deming's P10 is *Eliminate slogans, exhortations, and targets*. No slogan or exhortation ever helped a person to do a better job. High quality organizations frequently have slogans or commonly used exhortations but these come after the fact. Once the quality is there, slogans arise naturally. Until quality is visibly improving under a sincere improvement program, slogans have a negative effect. They are viewed as blaming the worker for low quality.

Eliminate Performance, "Merit", and Annual Reviews (DD3). They encourage short term thinking and discourage long term efforts and teamwork. People end up working for themselves rather than the organization. Performance reviews are discouraging – people lose time recovering from them. Typically, reviews are based on easily measured *random* numbers and they do not measure people's real value.

Rewarding merit is fine, *if that is what you are really doing*. True merit involves working *above* the capabilities of the system. It is a very rare event. The typical merit program rewards people on a random basis; this is counter productive. Within any system, performance varies. Purely by chance, some people perform better than others *within* the capabilities of the system. Randomly picking a tenth, or a quarter, or a half of your people to reward as meritorious can do nothing but discourage the other, equally hard working, people.

To find out if someone is truly working above the capabilities of the system, you need to know the capabilities of the system. This requires data and statistical analysis of the data to identify the system's capability. You should seek to find out what a person who works above the system's capability does differently. Perhaps the person seeks out better raw materials to work with. If you randomly identify people as meritorious, learning what they do differently is a waste of time and effort and discourages the search for quality. Similarly, seeking out the *particular* causes of defects that are built into the process is a waste of time.

Break Down Barriers to Communication (Essentially P9). Get people talking. In manufacturing concerns, purchasers, design, engineering, production, and sales people all need to communicate. All are part of the process, so all need to work together to improve the process. Moreover, suppliers and customers need to be involved in process improvement. Suppliers who do not know your needs cannot fill them. Similarly, you are the supplier to your customers.

1.1.3 Train, Retrain, and Educate

The key to higher quality, higher productivity, and lower costs is to *work smarter not harder*. Management's primary job is to provide workers the tools (intellectual and physical) to do this. Given the chance, innovative workers will actually invent most of the tools. Management's role is to identify good tools and put them to use. Working smarter requires training and education. These points are essentially Deming's P6 and P13.

Train people to perform their job. Teach them what their job is and how to do it. Teach them when the job is finished and whether it was done correctly. The best efforts of workers are futile if they do not know what to do. Deming (1986) gives example after example of people who were never taught what their job was or how to do it. They learned their jobs from other workers who had never been taught their jobs either. Motivating workers requires showing them how their job fits into the larger scheme of things. Money is a poor long term motivator.

When the process changes, the job changes. Retraining for the new job is required. Occasionally, people are found to be unsuited for a job, perhaps because of poor initial training. It is almost impossible to undo bad training. These people must be retrained for a new job. Retraining is part of driving out fear. Retraining allows workers to believe in the security of their jobs.

In addition to job training, management should assist in the general education of its employees. Education gives workers perspective on their jobs and their industry. A narrow view loses the opportunity of taking useful and creative contributions from other, not obviously related, fields. Working smarter rather than harder requires education. People's best efforts are not good enough. They must learn how to work smarter.

1.1.4 Insist on Action

Talk is cheap. Only action will improve quality. Start with little steps. Don't jump in with both feet. Keep it simple; keep it small. Begin by finding a process that is ripe for improvement, something where the results will be immediate and obvious. Juran and Gryna (1993) suggest that initial projects should last no longer than six months. Immediate and obvious results help convince workers, lower, and middle management that top management is serious about quality improvement. Stick with it! Build on a first success towards many others. There are many highly useful tools in developing a program for quality, e.g., Quality Control Circles, Statistical Charts, and Statistical Design of Experiments. However, without constancy of purpose and continued *action*, these tools are nothing more than management fads. *Insistence on Action* is essentially Deming's P14.

Improving the *process* of production and service can never stop. Always *base actions on good data and sound statistical analysis*.

1.2 Some Technical Matters

In addition to the general principles discussed in the previous section, there are some specific business practices on which Deming had strong opinions.

Stop Awarding Business on Price Tag Alone (P4). Every product and service requires raw materials to work with. If you have poor quality raw materials, the quality you produce suffers. Awarding contracts to the lowest bidder is a guarantee of getting low quality materials. Typically, you need to *work with one supplier* for a given input to ensure that the input is of appropriately high quality. Work with your suppliers to get what you need. It is hard to obtain quality materials from one supplier; it is virtually impossible with several suppliers.

Maintain your infrastructure. It is difficult to achieve quality in the face of frequent and random breakdowns of vital equipment. It is much better to maintain equipment on a regular schedule so that down time is planned and accounted for and so the equipment works when it is needed. Deming tells a story of a worker who told his supervisor about a bearing going out on a vital machine. The bearing could be replaced easily in a few hours. The supervisor, under pressure to meet his quota,

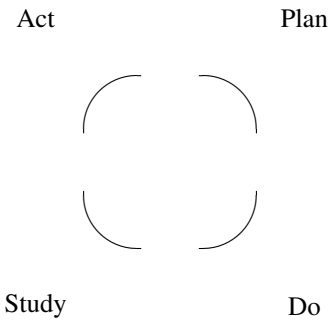

Figure 1.1: *Shewhart Cycle*

insisted that the worker continue using the machine. Inevitably, the bearing went out, causing major damage to the machine and much more extensive delays. As the old saw goes, "There is never time to do the job right but there is always time to do the job over." Fear is counter-productive.

Maintenance also applies to the most important part of the infrastructure: people. Obviously people are subject to "breakdowns" that impede their performance. Try to minimize these.

Cease Mass Inspection of Products and Services (P3). Quality is built into products and services. Once a product is made or a service performed, it is too late. Quality products and services do not require inspection. If the built-in quality is insufficient, inspect every unit to check whether it meets standards. Even this will not find all defective items. Note that producing defective products and services *costs more than producing quality products and services* because you pay once to produce them and again to repair them. When you buy poor quality, even if the producer makes good on defectives, the costs of producing and repairing defectives are built into the price you pay.

Deming mentions two other deadly diseases that apply in America: (DD6) *Excessive Medical Costs* and (DD7) *Excessive Costs Due to Litigation*.

1.3 The Shewhart Cycle: PDSA

A useful tool in improving quality is the Shewhart Cycle. It is a simple algorithm for improving quality. It can be applied almost anywhere.

1. Examine the process. Examine how can it be improved. What data are needed? Do they already exist? To study the effect of a change in the process, you generally need to change it.
2. Find existing data or conduct a (small scale) experiment to collect data.
3. Analyze the data. Plotting data or looking at tables may be sufficient.
4. Act on the results.
5. Repeat the cycle.

To put it briefly, Plan the investigation, Do the investigation, Study (or Check) the results, Act: Plan, Do, Study, Act: PDSA. The virtue of the Shewhart cycle is simply that it focuses attention on the key issues: using prior knowledge to evaluate the situation, collecting hard data and analyzing that data to evaluate the exact situation, and then acting on the analysis. Action is crucial, everything is a waste of time and resources if it does not result in appropriate action.

The US Air Force uses a similar Observe, Orient, Decide, Act loop.

1.4 Benchmarking

Another commonly used method of improving processes is *benchmarking*. Benchmarking consists of identifying the best processes and comparing yourself to the best.

EXAMPLE 1.4.1. Holmes and Ballance (1994) discuss the benchmarking efforts of a supplier. The supplier selected a world class partner and studied the processes of the partner. Some of the results they found are given below:

System	Supplier	Partner
Leadtime	150 days	8 days
Order input times	6 minutes	0 minutes
Late deliveries	33%	2%
Shortages per year	400	4.

This chart clearly shows how far short the supplier's performance fell relative to the partner. The supplier looks pretty awful, at least in these measures, when the partner's performance is used as a benchmark. But the chart misses the real point. Identifying how badly you are doing is only of value if it spurs improvement. The more valuable part of benchmarking involves examining, in this case, the partner's leadtime, order input, delivery, and inventory systems in an effort to improve the supplier's processes. Remember that *fixing blame does not fix problems*.

1.5 Exercises

EXERCISE 1.5.1. Watch the historically important *NBC White Paper* documentary on quality *If Japan Can, Why Can't We?* https://www.youtube.com/watch?v=vcG_Pmt_Ny4 When it comes to quality management everyone has a dog in the fight. I personally really like Deming's ideas. Banks (1993) takes a very different view. You should not accept anyone's pontifications blindly.

EXERCISE 1.5.2. Watch Deming's famous red beads experiment https://www.youtube.com/watch?v=ckBfbvOXDvU

EXERCISE 1.5.3. Watch the famous "funnel experiment" on why you should not tamper with a process that is under control. https://www.youtube.com/watch?v=cgGC-FPgPlA

EXERCISE 1.5.4. Watch Steve Jobs on Joseph Juran. https://www.youtube.com/watch?v=XbkMcvnNq3g Jobs was a cofounder of Apple, founder of Next, and CEO of both. I include this because I am a much bigger fan of Deming than of Juran so I thought Juran should get some time from someone who is a fan.

EXERCISE 1.5.5. Watch ASA's 2019 JSM *Deming Lecture* by Nick Fisher, "Walking with Giants." https://ww2.amstat.org/meetings/jsm/2019/webcasts/index.cfm?utm_source=informz&utm_medium=email&utm_campaign=asa&_zs=HpXOe1&_zl=HDD56#deming. There are other lectures on the link that you are not required to watch. (American Statistical Association) (Joint Statistical Meetings)

EXERCISE 1.5.6. Watch Scott Berry, "The Billion Dollar Statistical Concept" https://www.youtube.com/watch?v=XzenJPwZE_I. Valuable video but not of particular relevance to this class.

EXERCISE 1.5.7. Watch my dissertation advisor Donald A. Berry, "Multiplicities: Big Data = Big Problems" https://www.youtube.com/watch?v=ICOiKThwjoc. Valuable video but not of particular relevance to this class.

EXERCISE 1.5.8. Watch PBS's *Command and Control* http://www.pbs.org/wgbh/americanexperience/films/command-and-control/player/. This is a interesting documentary about a serious accident in a missile silo. The exercise is to write a summary of what is being done poorly and what is being done well in these systems. There may be issues with seeing the program.

EXERCISE 1.5.9. Watch NOVA's *Why Trains Crash* http://www.pbs.org/wgbh/nova/tech/why-trains-crash.html The exercise is to write a summary of what is being done poorly and what is being done well in the systems discussed. There may be issues with seeing the program.

Chapter 2

Basic Tools

Statistics is data analysis — in any form. Statistics is the science and art of making sense out of collections of numbers. Statistics can be as simple as graphing data or computing a few numerical summaries. It can be as complicated as developing complex statistical models and validating the assumptions underlying those models. Conclusions from statistical analysis can be as simple as stating the obvious (or rather what has become obvious after graphing and summarizing), or conclusions can be formal results of statistical tests, confidence intervals, and prediction intervals. In the Bayesian approach, conclusions take the form of probability statements about the true condition of the process under consideration.

Underlying all modern statistical procedures is an appreciation for the variability inherent in data. Often, appreciating that data involve variability is referred to as "statistical thinking." The temptation is to over interpret data. To think that what occurred today is a meaningful pattern, rather than the randomness that is built into the system. As the variability in the data increases, there is more need to use formal statistical analysis to determine appropriate conclusions.

At the other end of the spectrum, statistics are needed to summarize large amounts of data into forms that can be assimilated. Statistics must study both how to appropriately summarize data and how to present data so that it can be properly assimilated.

An important aspect of data analysis is making comparisons: either to long time standards (known populations) or to other collected data. As with other statistical procedures, these comparisons can be made either informally or formally.

To know the current state of a process or to evaluate possible improvements, data must be collected and analyzed. At its most sophisticated, data collection involves sample surveys and designed experiments. The only way to be really sure of what happens when a process is changed is to design an experiment in which the process is actually changed. Sometimes it is more cost effective to use data that are already available. In any case, sophisticated data typically require a sophisticated analysis.

Often, great progress can be made collecting simple data and using simple statistical techniques. Simple charts can often show at a glance the important features of the data. *Histograms* are a good way of showing the main features of a single large set of data. Three other charts that are useful, if not particularly statistical, are *Cause and Effect diagrams*, *Flow charts*, and *Pareto charts*. All of these charts are discussed in this chapter. Control charts are used to evaluate whether processes are under statistical control, cf. Chapter 4. *Scatter plots* show the relationship between pairs of variables, cf. Chapter 5. *Run charts* are simply scatter plots in which one of the two variables is a time measurement, cf. Chapters 4 and 6. The points in a run plot are often connected with a solid line; the temporal ordering of the points makes this reasonable.

2.1 Data Collection

On one hand, data are worthless without a proper analysis. It is amazing how often people who spend large amounts of time and money on collecting data think that they need to put almost no resources into properly analyzing the data they have so painstakingly collected. On the other hand,

no amount of data analysis can give meaning to poorly collected data. A crucial aspect of statistical data analysis is proper data collection.

Data need to be germane to the issue being studied. Plans must be made on how the data collected can and will be used. Collecting data that cannot be used is a waste of time and money. Data should not be collected just because they are easy to collect. Collecting data on the number of times a person hits the "k" key on their computer keyboard is probably worthless. We suspect that collecting data on the number of keys hit during the day is also probably worthless. Numbers such as these are typically used to make sure that workers are working. Obviously, smart workers can find ways to beat the system. But more importantly, if management has no better idea of what is going on in the office than the number of keystrokes workers hit in a day, they have much more profound problems than workers loafing. Data should be collected because they give information on the issues at hand. It is a sad state of affairs when the best data that management can come up with on the condition of their workplace is how often employees hit their keyboard.

At the beginning of the third millennium AD (or CE), one of the greatest changes in society is the ease with which some types of data can be collected. Traditionally, data collection has been very difficult. Along with the new found ability to collect masses of cheap data have come techniques that try to separate the data wheat from the data chaff. What constitutes wheat changes from problem to problem and there is no guarantee that an easily collected set of data will contain any wheat. This is an apt time to recall that Deming's fifth deadly disease (DD5) is essentially running a company on easily collected data alone.

It is a capital mistake to act on data that give an incomplete picture of the situation. Transferring the computer salesperson who has the lowest monthly sales (easily collected data) will be a disaster if that person has informally become the technical resource person that all the other sales people need in order to make their sales.

As discussed in Chapter 1, a key part of the Shewhart cycle for quality control and improvement is the collection of appropriate data. There are many types of data that appear in business applications. Some common types of data are

Process control data: data that are used to establish that an industrial (or service) process is under control and thus that reliable predictions can be made on the future of this process.

On-line control data: data used to fine tune industrial processes. A key feature in on-line control is the need to not overcontrol a process. A process that is under control should not be tampered with. "Fine tuning" a process that is undercontrol actually decreases quality because it adds to the variability in the process, cf. Deming (1986, p. 327).

Inspection data: data that are used to decide whether a batch of goods are of sufficiently high quality to be used in production or to be shipped as products to customers. As alluded to in Chapter 1 and as will also be discussed in Chapter 8, a major goal is to put an end to inspection.

Observational data: data that are collected on the current state of affairs. Control data are observational, but more generally, observational data can be taken on several variables and used to suggest relationships and give ideas for solutions and/or experiments. (Easily collected electronic data tend to fall in this category.)

Experimental data: data that are obtained from a formal experiment. These are the only data that can be reliably used to determine cause and effect. This is discussed in Chapter 9. Experiments are generally used for product improvement and for isolating the causes of major problems, i.e., problems that are not getting solved by other means.

Data collection should lead to action. Good data on specific issues are typically expensive to collect and they should be collected for a reason. The reason for collecting data is that the data can lead to useful action. In this day and age, collecting observational data is often very easy. Electronic devices can collect huge masses of data: data that may never get examined and data that may contain very little useful information. If such data are inexpensive to collect and store, then it might be of some marginal value to do so, on the off chance that at some point in the future they might have

Table 2.1: *Causes of unplanned reactor shutdowns.*

Cause	Frequency	Percentage
Hot melt system	65	38
Initiator system	25	15
Cylinder changes	21	12
Interlock malfunction	19	11
Human error	16	9
Other	23	14
Total	169	100

some value. But such data should not be collected and stored with the expectation that they are useful merely because they are readily available. (This is quite distinct from the issue of trying to find uses for the data that are available!)

In fact, while the Shewhart cycle illustrates the need to collect data, it can also be used as an algorithm for proper data collection.

Plan the process of data collection. How to collect the data. What to collect. How to record it. How to analyze it. Often the analysis is hampered by recording the data inappropriately!

Collect the data. Do it!

Analyze the data. Study it and learn from it. Good data are often expensive to collect; resources have to be put into learning as much as possible from the data. Unanalyzed data, improperly analyzed data, and poorly analyzed data are all a waste of time, effort, and money.

Take action based on the results of the data analysis. The data should have been collected for a reason. Address that reason. Often, back at the planning stage, one can set up contingencies indicating that if the data come out like this, our actions will be these. (But it is unwise to completely tie oneself to such plans, because the analysis of the data may indicate new options that did not appear at the planning stage.)

2.2 Pareto and Other Charts

In this section we illustrate Pareto charts and other charts including bar charts and pie charts. Pareto charts are simply bar charts that are arranged to emphasize the *Pareto principle* which is that in any problem, a few factors are responsible for the bulk of the problem. The importance of Pareto charts and other charts is simply that they convey information very rapidly and make a visual impact. Of course, it is important that the visual impact made be an accurate representation of the data being portrayed.

EXAMPLE 2.2.1. Juran and Gryna (1993) present historical data on the causes of unplanned reactor shutdowns. These are given in Table 2.1. Note that the table has been arranged in a particular order. The largest cause of shutdowns is listed first, the second largest cause is listed second, etc. The only exception is that the catch-all category "other" is always listed last. A Pareto chart is simply a bar chart that adheres to this convention of the most important cause going first. A Pareto chart is given in Figure 2.1. The vertical scale on the left of the diagram gives the raw frequencies while the vertical scale on the right gives percentages. The line printed along the top of the diagram gives cumulative percentages, thus the first two categories together account for 53% of unplanned shutdowns and the first three categories together account for 65% of shutdowns.

Compare your immediate reactions to Table 2.1 and Figure 2.1. Don't you find that the information is presented much more effectively in Figure 2.1?

The point of Figure 2.1 is to illustrate that the main cause of shutdowns is the hot melt system. In order to reduce the number of shutdowns, the first order of business is to improve the performance of the hot melt system. It is interesting to note that prior to collecting this historical data, the

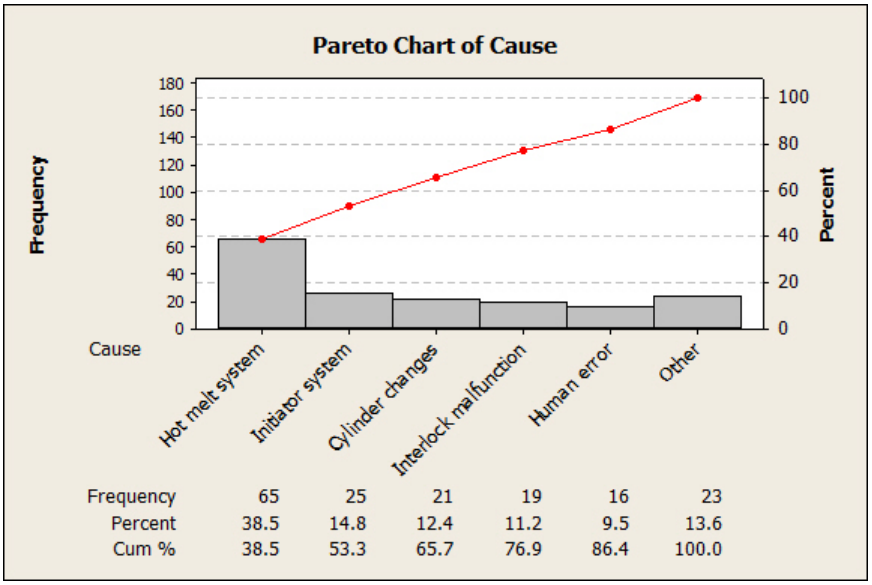

Figure 2.1: *Pareto chart: causes of unplanned reactor shutdowns.*

Table 2.2: *Costs of unplanned reactor shutdowns by cause.*

Cause	Frequency	Cost per Shutdown	Total Cost	Percentage of Cost
Hot melt system	65	1	65	26
Initiator system	25	3	75	30
Cylinder changes	21	1	21	8
Interlock malfunction	19	2	38	15
Human error	16	2	32	13
Other	23	1	23	9
Total	169		254	100

reactor personnel thought that cylinder changes would be the primary cause of unplanned reactor shutdowns.

Any cause in the Pareto chart can be further broken down into its constituent parts and a new Pareto chart formed for that cause. Typically, one would do this for the most import cause, but that would get us into the details of the hot melt system. However, we can illustrate the same idea more accessibly by breaking down the least important category. (Since it is the least important, it will matter least if our speculations are jeered at by the nuclear engineering community.) Perhaps the 16 shutdowns due to human error could be broken down as: inadequate training, 11; asleep at the switch (otherwise known as the Homer Simpson cause), 3; other, 2. Clearly a Pareto chart can be formed from these.

An alternative to a Pareto chart based on frequencies of shutdowns is a Pareto chart based on costs of shutdowns. Often when a problem occurs, say a manufacturing process shuts down, the cost of fixing the problem depends on the cause of the problem. It may be more important to decrease the cost of shutdowns rather than the number of shutdowns. Table 2.2 incorporates costs into the causes of reactor shutdowns. Note that this is no longer a Pareto table because the most important cause (as measured by cost now) is no longer listed first. Figure 2.2 gives a Pareto chart for shutdown causes as ranked by cost.

Figure 2.3 gives a pie chart of the frequencies of unplanned reactor shutdowns. Figure 2.4 gives a pie chart of the costs of unplanned reator shutdowns

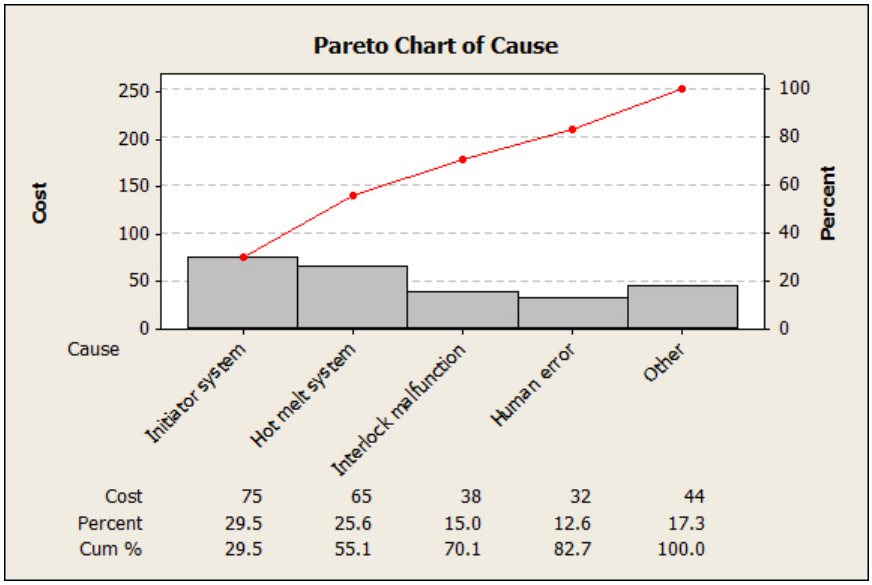

Figure 2.2: *Pareto chart: causes of unplanned reactor shutdowns ranked by cost.*

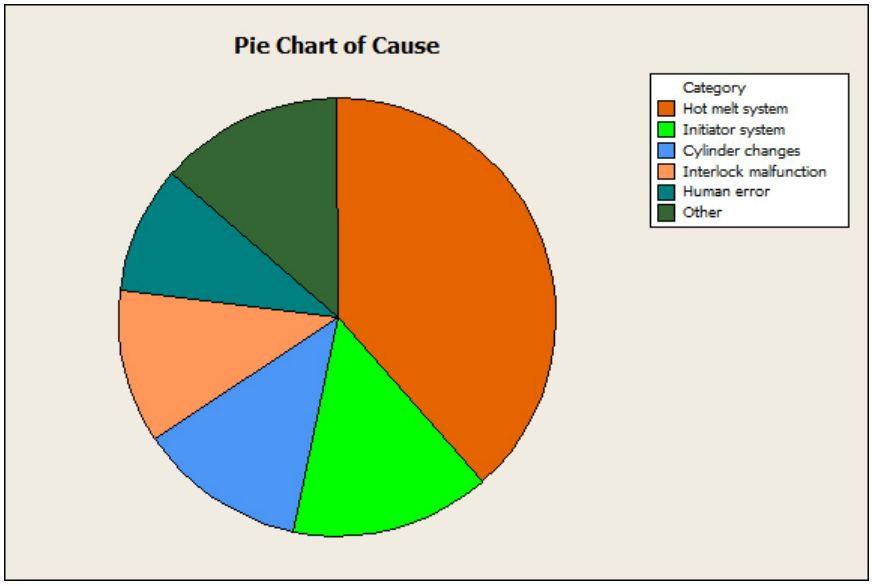

Figure 2.3: *Pie chart: causes of unplanned reactor shutdowns ranked by frequency.*

2.3 Histograms

Data have variability. An easy way to display variability is through a histogram. A histogram is just a bar chart that displays either the frequencies or relative frequencies of different data outcomes.

EXAMPLE 2.3.1. Figure 2.5 gives a histogram for the heights, as measured in inches, of 781 people who rode a roller coaster one day. The most frequently observed height is 66 inches. This is referred to as the mode. Most of the observations fall between 63 and 71 inches (85%).

Note the sharp drop off at 62 inches. This histogram is skewed, rather than symmetric, because of the sharp drop off. This suggests that people under 62 inches are probably not allowed to ride this

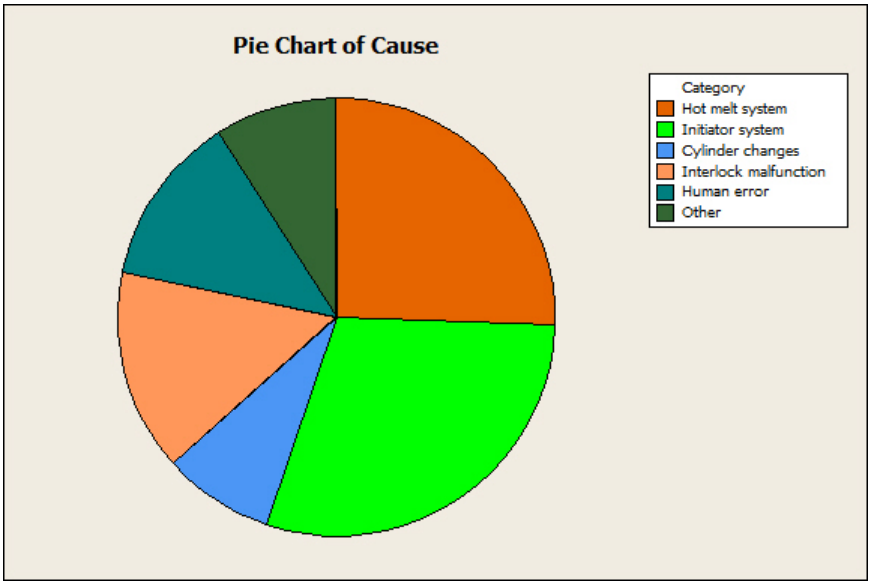

Figure 2.4: *Pie chart: causes of unplanned reactor shutdowns ranked by cost.*

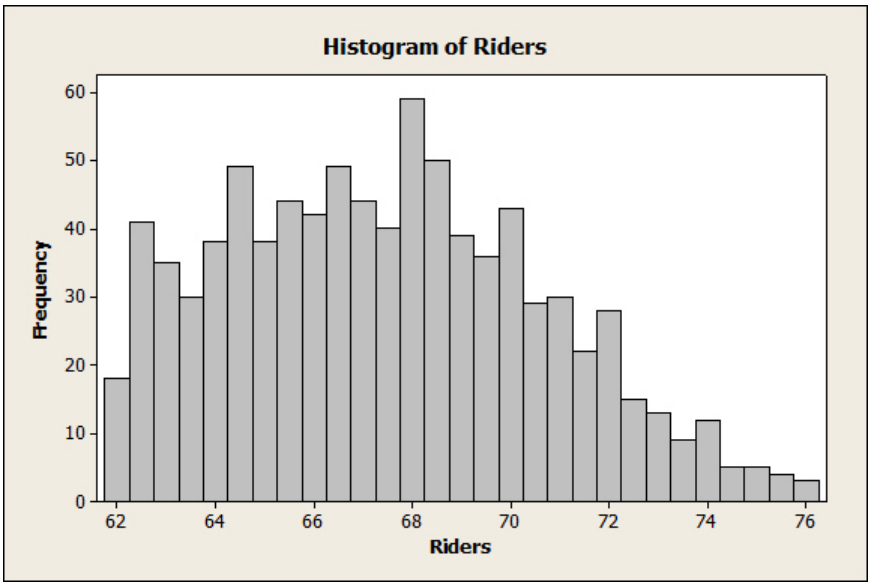

Figure 2.5: *Histogram of 870 rollercoaster riders.*

particular roller coaster. In industrial applications, a histogram that looks like this would suggest just about the same thing, i.e., that someone has inspected all of the material to make sure that it satisfies a given specification. A sharp drop off at the top end would suggest that items have also been inspected to see if they are too large.

There are a couple of problems with such inspection schemes. Probably the less significant problem is that some "defective" items will pass the inspection. Some people under 62 inches will actually be allowed to ride the roller coaster. A more serious problem occurs when we buy items that are required to meet specifications. (Dare we think of wanting to buy human beings that are supposed to be at least 62 inches tall?) When we buy items from a manufacturer who is producing

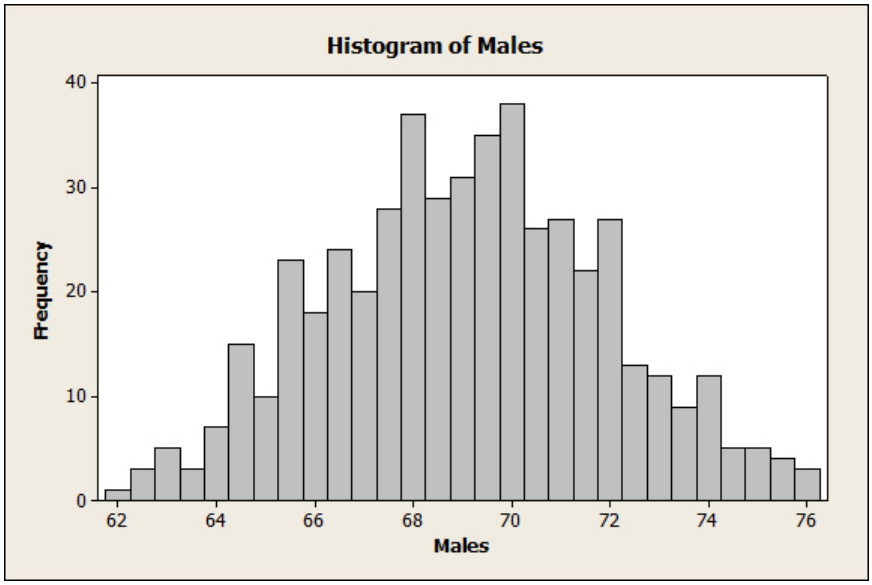

Figure 2.6: *Histogram of 492 males.*

defective items along with the good items, *even if we never explicitly buy a defective item*, the cost of the good items we buy have the overhead cost of producing the defective items built into them. If the manufacturer improves the process so that it produces only acceptable items, the cost of those items should go down!

For the height data, it is obvious that Figure 2.5 is displaying data that are really a combination of two subgroups having different characteristics: males and females. Figure 2.6 gives the histogram of roller coaster rider heights for males. Note that the histogram for males has a nice symmetric shape with a mode of 67 inches. There is no sharp cutoff at 62 inches, because the heights naturally decrease near 62 inches. Most heights are between 66 and 71 inches and almost all heights are between 63 inches and 75 inches.

If we think of Figure 2.6 as showing results from an industrial process that is designed to make products that are at least 62 units long, this process is not doing too badly. Almost all of the items produced are greater than 62 and not many are produced that are even close to 62.

Figure 2.7 gives the histogram for females. This histogram again shows a sharp cutoff at 62 inches, causing it to appear skewed rather than symmetric. Even though there is not overt discrimination here against women, the cutoff value is set at a height that clearly causes implicit discrimination against females. One must weigh the safety concerns that motivated the creation of this cutoff value against this implicit discrimination.

If we think of Figure 2.7 as showing results from an industrial process that is designed to make products that are at least 62 units long, this process is not doing very well. The sharp drop off at 62 suggests that we are not seeing all of the production; that we are seeing only those units that exceed the specification. The sharp drop off at 62 suggests that even though we will not get defective items sold to us, we will be paying for producing defective items as part of the overhead involved in buying good units.

Finally, Figure 2.8 gives a histogram with an interesting shape. The histogram has more than one peak, which is referred to as being multimodal. The histogram was arrived at by combining data from 100 female softball players, most of whom have heights around 65 to 69 inches, and male basketball players, whose heights tend to center around 78 inches. The moral is that multimodal histograms are often caused by combining groups that perhaps should not be combined. Note how-

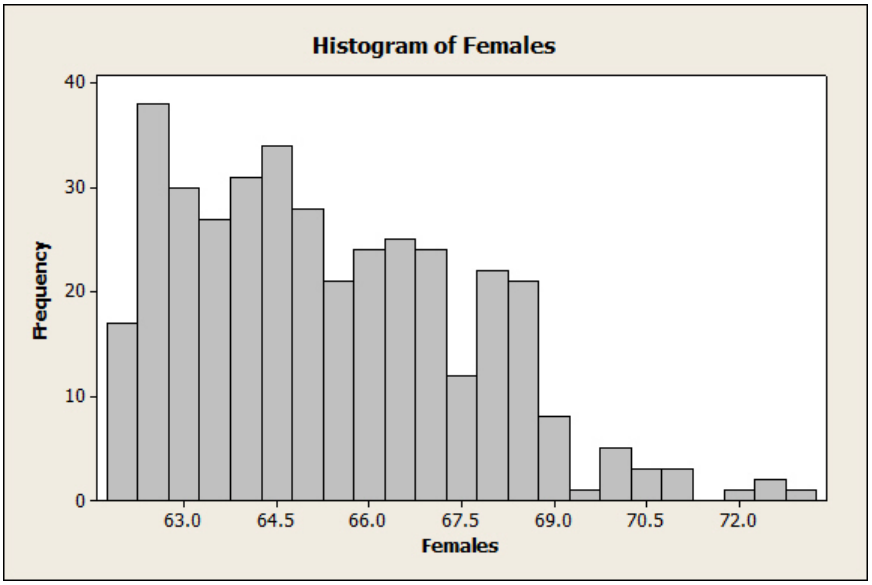

Figure 2.7: *Histogram of 378 females.*

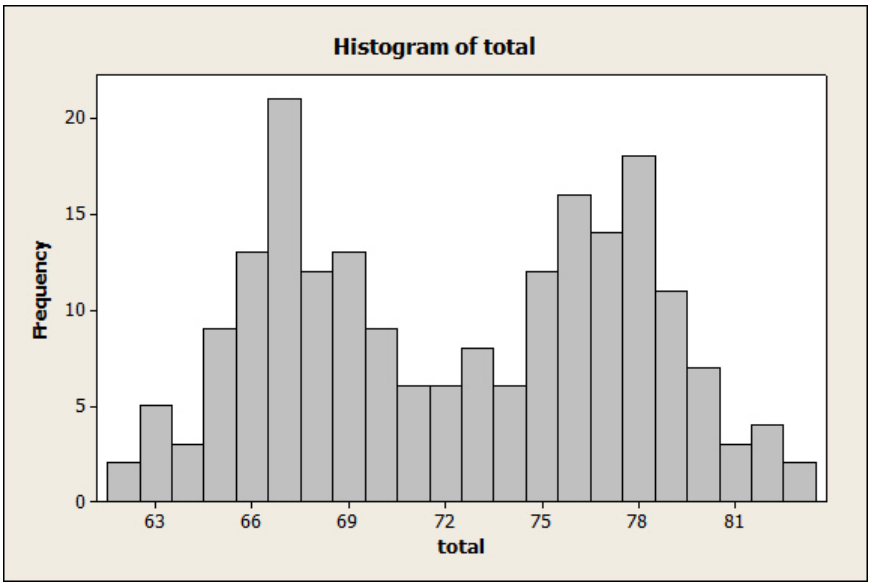

Figure 2.8: *Histogram of heights.*

ever that combining groups does not always create a multimodal histogram as we saw in Figures 2.5 through 2.7.

2.3.1 Stem and Leaf Displays

Stem and leaf displays provide informal histograms yet give most of the original data. They are very easy to construct and they lose almost no information in the data.

EXAMPLE 2.3.2 *Vinyl Floor Covering Data.*

Table 2.3: *Tear Test Results from Three Vinyl Floor Coverings - qe98-284*

Case	A	B	C
1	2288	2592	2112
2	2368	2512	2384
3	2528	2576	2096
4	2144	2176	2240
5	2160	2304	2320
6	2384	2384	2224
7	2304	2432	2224
8	2240	2112	2368
9	2208	2288	2112
10	2112	2752	2144

Depth	Stem	Leafs
3	21	146
3	22	048
4	23	068
1	24	
1	25	2

Figure 2.9: *Stem and Leaf Display of Floor Covering A. (Leafs are 10s.)*

Table 2.3 gives data on tear test values for three vinyl floor coverings from Phillips et al. (1998). We will use these data to illustrate stem and leaf displays in this subsection, dot plots in the next subsection, and box plots in the next section.

Figure 2.9 gives the stem and leave display for the tear scores of vinyl type A. The display has three parts, depth, stems, and leafs. The scores are 4 digit numbers. The last digit is ignored, the first two digits are used as the stems and a histogram is created using the third digit. The third digits are called the leaves. Thus in the first row of the display, the stem is 21 and the leafs are 1, 4, and 6. Using an x to indicate that we do not know the fourth digits, the first row indicates that the numbers 211x, 214x, and 216x are tear scores for vinyl A. From the second row, the numbers are 220x, 224x, and 228x. Thus, except for the fourth digits, the stem and leaf display is giving all of the data, while providing a histogram (displayed on it's side). In particular, at a glance we see that there are equal numbers of observations in the 21xx's the 22xx's, and the 23xx's. It should be noted that stem and leaf displays are very easy to construct by hand. For small or moderate sized collections of data, construction of a stem and leaf display should often be the first statistical analysis.

The depth column for the stem and leaf display is designed to help in finding the median observation, i.e., the observation that has as many values greater than it as there are numbers less than it. The depth starts at both ends of the display, and counts the number of observations from the top or bottom of the display. Thus, from the bottom, there is one number in the 25xx's, cumulatively, one number among the 25xx's and 24xx's, and a total of four numbers among the 25xx's, 24xx's, and 23xx's. Counting from the bottom, we stop before the 22xx's because including them would include more than half the data – in this case there are 10 observations, so 5 observations would give half of the data. Starting from the top down, the 21xx's have three observations, but including the 22xx's would get above 5, so they are not included. The number of observations in the 22xx's is displayed separately. To find the median, the 5 smallest observations are 224x and smaller, and the 5 largest observations are 228x and larger, so the median would be taken to be the average of 224x and 228x.

Figure 2.10 provides an alternative stem and leaf display in which the stems have been broken in half. Thus the first stem includes any numbers from 210x to 214x and the second stem includes numbers from 215x to 219x. This gives a more detailed histogram.

Finally, we can compare two sets of numbers by constructing back to back stem and leaf displays. This is illustrated for vinyls A and B in Figure 2.11. The stems are placed in the middle of the display, the leafs for vinyl B are given to the left of the stems and the leafs for A are given to

Depth	Stem	Leafs
2	21	14
3	21	6
5	22	04
5	22	8
4	23	0
3	23	68
1	24	
1	24	
1	25	2

Figure 2.10: *Alternative Stem and Leaf Display of Floor Covering A. (Leafs are 10s.)*

B	Stem	A
71	21	146
8	22	048
80	23	068
3	24	
971	25	2
	26	
5	27	

Figure 2.11: *Back to Back Stem and Leaf Displays for Vinyls B and A of Floor Covering A. (Leafs are 10s.)*

the right of the stems. It is clear that vinyl B tends to have larger tear results than vinyl A. However, the minimum tear results are nearly the same for both vinyls. Note that back to back stem and leaf displays only allow comparison of two vinyls, and we have three vinyls in the data.

2.3.2 Dot Plots

Dot plots provide yet another form of histogram. They consist of a plot of an axis (scaled to be appropriate for the data) along with dots above the line to indicate each data point. Figure 2.12 gives the dot plot for the tear test results for vinyl A.

Dot plots can be used to give a visual comparison of several groups of observations. Figure 2.12 gives dot plots on a common scale for all three vinyls in Table 2.3. We can see that the observations for vinyl B tend to be a bit larger and perhaps more spread out than those for vinyls A and C. Vinyls A and C look quite comparable except for one observation just over 2520 in vinyl A.

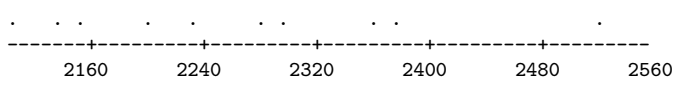

Figure 2.12: *Dot Plot of Floor Covering A.*

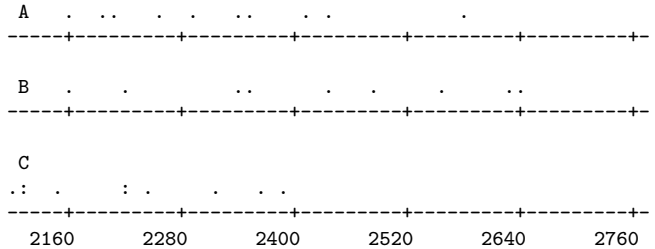

Figure 2.13: *Dot Plots for Three Floor Coverings.*

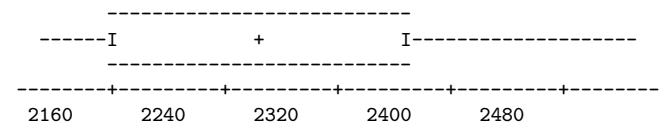

Figure 2.14: *Box Plot of Floor Covering A.*

Table 2.4: *Summary Statistics*

Vinyl Cover	MIN	Q1	MEDIAN	Q3	MAX
A	2112	2156	2264	2372	2528
B	2112	2260	2408	2580	2752
C	2096	2112	2224	2332	2384

2.4 Box Plots

Box plots give an alternative visual display of the data based on a five number summary of the data.

EXAMPLE 2.4.1 *Vinyl Floor Covering Data.*
Figure 2.13 gives the box plot for the vinyl A data of Table 2.3. Note that the overall impression of the plot is one of symmetry. The mark in the middle of the box is near the center of the box, and the lines on the edges are not too dissimilar in length.

The five numbers on which box plots are based are the maximum observation, the minimum observation, the median (the point that has half the data smaller than it and half the data larger), the first quartile Q1 (the point that has $1/4$ of the data smaller than it and $3/4$ of the data larger), and the third quartile Q3 (the point that has $3/4$ of the data smaller than it and $1/4$ of the data larger). The plot makes a box using Q1 and Q3 as ends of the box, marks the median inside the the box, and creates "whiskers" from the ends of the box out to the maximum and minimum. Many programs for computing box plots define inner and outer fences and identify two classes of outliers as points that fall outside the inner and outer fences.

EXAMPLE 2.4.1 *Vinyl Floor Covering Data continued.*
Table 2.4 gives the five summary statistics on which the box plots for the vinyls are created. The median for cover A is actually any number between 2240 and 2288 but it is reported as the mean of these two numbers.

To compare the vinyls visually, we can plot box plots for all three on a common scale. Figure 2.14 reconfirms the impression that vinyls A and C are roughly similar, with vinyl C looking a little smaller and more compact. Vinyl B tends to be larger and more spread out but having the same lowest observations.

While the length of the box plot is $Q3 - Q1$, when plotting multiple box plots with very different sample sizes it can be useful to use the square root of the sample size as the width of the box plot. McGill, Tukey, and Larsen, (1978) discuss this and other variations on box plots.

2.5 Cause and Effect Diagrams

Cause and Effect diagrams, also known as Fishbone and Ishikawa diagrams, are a useful quality improvment tool but are not really a statistical tool in that they do not involve analyzing numerical data. Cause and Effect diagrams are just that, they diagram the potential causes of an effect. These potential causes can then be explored to see what is the real cause of an effect. Usually, the effects

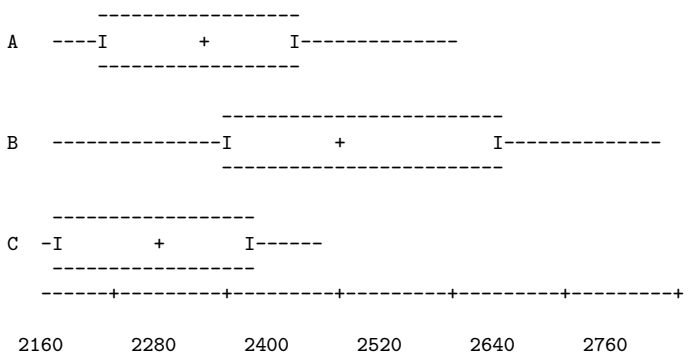

Figure 2.15: *Box Plots for Three Floor Coverings.*

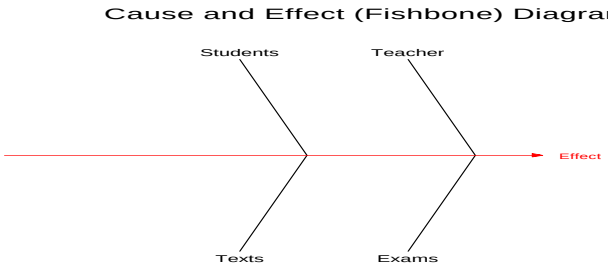

Figure 2.16: *Cause and Effect diagram with only primary causes.*

of interest are somehow undesirable, the purpose of the cause and effect diagram is to brainstorm possibilities for fixing problems.

EXAMPLE 2.5.1 Figures 2 .15 and 2.16 illustrate cause and effect diagrams for a statistics class. At the end of the spine, is the effect of interest, the statistics class. On the primary ribs coming off of the spine are the Primary causes. In Figure 2.15 these are listed as the Students, the Teacher, The Text books, and the Exams. If Figure 2.16 additional detail is given to each primary cause, identifying secondary causes. For Students these are the course prerequisites, their home life, and their study habits. For the Teacher, these are the teacher's knowledge, interest and communication skills. Figure 2.17 presents the general idea.

2.6 Flow Charts

Flow charts (also known as *process maps*) are used to graphically display the steps in a process.

Minitab has support for constructing flow charts, see
support.minitab.com/en-us/workspace/get-started/map-your-process/

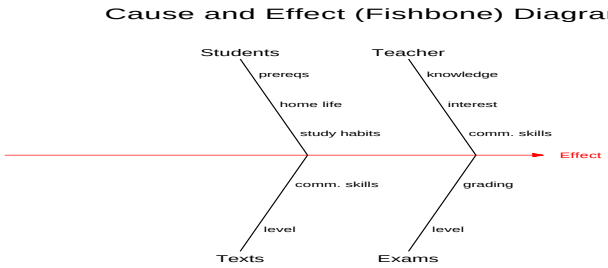

Figure 2.17: *Fishbone diagram with both primary and secondary causes.*

Flow charts can be made in R.

A sophisticated package that allows easy construction is `DiagrammeR`. After loading the package the following simple code illustrates three simple flow charts.

```
library(DiagrammeR)
DiagrammeR("graph LR;
           A-->B;
           B-->C;
           B->>D")

DiagrammeR("graph TB;
    A(Rounded)-->B[Squared];
    B---C{Rhombus!};
    C-->D>flag shape];
    C-->E((Circle));")
DiagrammeR("graph TB;
    A(Both Working)-->B[One Working];
    A-->C{Both \n Failed};
    B-->A;")
```

Basic programing information is given at
https://cran.r-project.org/web/packages/DiagrammeR/DiagrammeR.pdf and the author provides a video at http://rich-iannone.github.io/DiagrammeR/. I found out about this from
https://sites.temple.edu/psmgis/2017/07/30/flow-charts-in-r-using-diagrammer/

An alternative is the R package `diagram`.

EXAMPLE 2.6.1 Figure 2.18

Chapter 3

Probability

3.1 Introduction

In this chapter we review the basic notions of probability. Simply put, probability models are used to explain random phenomena. Before we define a probability, let us try to understand how probabilities are used and how probabilities are calculated.

Imagine that you are driving to work and you are interested in how many people are driving the same model car as you. Your *experiment* consists of counting cars of the same model that you encounter on your drive to work. The *actual* outcome of this experiment cannot be predicted with certainty. The set of all *possible* outcomes is called the *sample space* and it is denoted by S. For this experiment, the sample space is $S = \{0, 1, 2, ..., 500\}$ where we assume that you encounter at most 500 cars on your drive to work. Within the sample space, we can define *events* of interest. For example, you may want to make sure that your's is a relatively exclusive model and you will believe this if you see only 0, 1, or 2 others like it. We would express this event as $A = \{0, 1, 2\}$. You may decide that the masses are driving this car if you see 3 or more cars of the same model on the road. This event is expressed as $B = \{3, 4, 5, 6, 7..., 500\}$.

Events are subsets of the sample space. Generally, we use capital letters from the beginning of the alphabet to denote events. In examples like this in which we can list the possible outcomes, probabilities are typically assigned to outcomes and are computed for events. Probabilities are always between 0 and 1 and can posibly be either 0 or 1. Something that happens with probability 1 is a sure thing and something that has no chance of happening has probability 0. Something must *always* occur, i.e. some outcome in the sample space must always occur because the sample space is a listing of everything that can occur. Therefore the probability of the sample space is 1. If some event occurs with a certain probability, say 0.3, then the probability that it will not occur is $1 - 0.3$. If two events are mutually exclusive, they cannot both happen at the same time. For example, on your drive to work, if you count a total of 3 or more cars, then the event $A = \{0, 1, 2\}$ cannot happen, because the event B has occurred.

The easiest way to compute probabilities are in situations where outcomes are equally likely. We are all familiar with probability models such as tossing coins and rolling dice. In situations where every outcome in the sample space is equally likely, we compute the probability of an event E as

$$\Pr(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in the sample space}}$$

This formula is a way of computing probabilities and is not a definition of probability. A rigorous definition of probability is quite difficult, as difficult as defining concepts such as force and velocity in physics.

We now examine an example in which the outcomes do not have equal probabilities.

EXAMPLE 3.1.1. Consider six outcomes having to do with your driving habits and the posibility of you getting a speeding ticket. The outcomes are all combinations of two factors, first whether you get a speeding ticket (Y) or do not get a speeding ticket (N) and second, whether you never drive over the speed limit (NDO), you sometimes drive over the speed limit (SDO), or you always drive

Table 3.1: *Speeding–Getting a Ticket probabilities*

		Speeding		
		NDO	SDO	ADO
Speeding	*Y*	0.0	0.18	0.30
Ticket	*N*	0.12	0.30	0.10

over the speed limit (*ADO*). The combinations of these factors define six possible outcomes. The probabilities are listed in Table 3.1.

Note that each of the probabilities is between 0 and 1 (inclusive) and that each outcome has a different probability. The probability of an event such as the event that you sometimes drive over (SDO) the speed limit is computed as

$$\begin{aligned} \Pr(\text{SDO}) &= \Pr[(\text{SDO}, Y) \text{ or } (SDO, N)] \\ &= \Pr(\text{SDO}, Y) + \Pr(\text{SDO}, N) \\ &= 0.30 + 0.18 \\ &= 0.48\,. \end{aligned}$$

Similarly, $\Pr(\text{NDO}) = 0.12$ and $\Pr(\text{ADO}) = 0.40$. These are called marginal probabilities (because typically they would be written in the margin of the table). The marginal probabilities for whether or not you get a ticket are

$$\Pr(Y) = 0.48, \quad \Pr(N) = 0.52.$$

These computations are illustrations of the *addition rule* for probabilities. If A and B are events that cannot happen simultaneously, e.g., you cannot simultaneously get a tecket and not get a ticket while sometimes driving over the speed limit, then $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$.

We can also compute *conditional probabilities*. These are probabilities of events that are conditional on the knowledge of some other event. In our example, suppose your beloved knows that you got a ticket on your way to work and he/she claims that you *always* drive over the speed limit. Now, to defend yourself, you want to compute the probability that you always drive over (ADO) the speed limit given that you received a ticket (Y). This is written $\Pr(\text{ADO}|Y)$ and read as "the probability of ADO given that the event Y has occurred". It is computed as the joint probability of the events ADO and Y divided by the marginal probability of the event Y,

$$Pr(\text{ADO}|Y) = \frac{\Pr(\text{ADO}, Y)}{\Pr(Y)} = \frac{0.3}{0.48} = 0.625.$$

Notice that the probability that you always drive over is $\Pr(\text{ADO}) = 0.4$ but given that you received a speeding ticket, the conditional probability has jumped to 0.625. This says that the occurrence of the event Y which tells your beloved (and anyone else) that you received a speeding ticket, has provided them with *additional information* about your driving habits.

We say two events, A and B are *independent* if knowledge of one event provides no information about the other. In this case, the probability that both events occur is the product of their marginal probabilities

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B).$$

In our example, we see that the event that you always drive over the speed limit is not independent of the event that you get a speeding ticket

$$\Pr(\text{ADO and } Y) = 0.3 \neq (0.4 \times 0.48) = \Pr(\text{ADO})\Pr(Y).$$

Another, more intuitive, way to check for independence is to see whether the conditional probability

of an event is altered from its unconditional probability, e.g., is $\Pr(\text{ADO}|Y) = \Pr(\text{ADO})$? If these probabilities are the same, we say the events are independent. In our example $Pr(\text{ADO}|Y) = 0.625$ and $Pr(\text{ADO}) = 0.4$, so kowing the event Y occurs provides additional information about ADO, in this case making ADO more probable. Generally, independence is a property that is assumed, not proven.

3.2 Something about counting

In order to define our sample space, we are interested in the number of ways in which something can happen. *Combinatorics* is the field that is concerned with counting. For example, suppose you are interested in the number of ways that you can get dressed in the morning. Suppose you're a poor college student and you own exactly 2 shirts that you plucked out of a dumpster, one is purple and the other is orange. You also own 3 pairs of pants that were given to you by various loving family members: their colors are fuschia, emerald, and poinsettia. Now, without regard to fashion or style, the number of ways that you can get dressed on this particular day is 6. For each shirt you choose, there are 3 possible pairs of pants to choose and so you have 2 shirts $\times$ 3 pants $=$ 6 possible outcomes. Notice that if you were limited by concern about untoward fashion statements, your number of ways to get dressed would be different.

Our shirts and pants example is an illustration of the *Multiplication Principle of Counting:* Suppose we can select an object X in m ways and an object Y in n way. Then the pair of objects (X,Y) can be selected in mn ways.

EXAMPLE 3.2.1. New Mexico car license plates used to have 3 digits followed by 3 letters. The total number of possible license plates without restriction is:

$$(10 \times 10 \times 10) \times (26 \times 26 \times 26).$$

There are 10 choices for each number and 26 choices for each letter. Of course, New Mexico car licence plates are limed by concern about untoward three letter verbal statements.

Suppose now that we are not allowed to repeat any number or letter. There are still 10 ways to pick the first number but once that is picked, there are only 9 ways to pick the second number. The number of possible license plates without repetition of a number or letter is:

$$(10 \times 9 \times 8) \times (26 \times 25 \times 24).$$

Here, the order in which the items are selected is important. The licence plate 123 ABC is different from the plate 213 CBA. An ordered arrangement of 3 similar items is called a *permutation*. We would not talk about a permutation in regard to a licence plate because it involves dissimilar items, both numbers and letters. We *would* talk about a permutation of the numbers (or the letters). Counting the number of licence plates without replications involves the $10 \times 9 \times 8$ different permutations of the ten numbers taken three at a time as well as the permutations of the $26 \times 25 \times 24$ different permutations of the twenty six letters taken three at a time.

Permutation: The number of permutations of n objects taken k at a time is

$$P_k^n = n(n-1)(n-2)\cdots(n-k+1).$$

A succinct way to write the right hand side of the above equation is

$$P_k^n = \frac{n!}{(n-k)!}$$

where $n! \equiv n(n-1)(n-2)\cdots(2)(1)$ and $0! \equiv 1$. Our answer for the number of possible New Mexico license plates without repeating a number or letter can be re-expressed in this notation as

$$P_3^{10} \times P_3^{26} = 11,232,000.$$

In New Mexico, there are plenty of licence plates for the number of cars, but one can see that in, say, California, problems might develop.

Often, we are interested in arrangements where order is not important. For example, suppose your nutritious lunch box contains an apple, a tangerine, and a pear. If you cared about what order you ate them in, you would eat lunch in $P_3^3 = 3!$ ways. From the point of view of your taste buds, it matters which fruit you eat first, second and third. However, from the point of view of your stomach, it merely gets a lump of all 3 fruit and therefore only one end result. This way of selecting 3 fruit from 3 fruit without regard to order is called a combination. If you had 3 fruits available, but promised a colleague first choice, so you only eat 2, there are three different lumps that could exist in your stomach: [apple, tangerine], [apple, pear], and [tangerine, pear]. Recall that to your stomach [tangerine, pear] is the same combination as [pear, tangerine].

Combination: An unordered arrangement of k objects selected from n objects is called a combination and is expressed as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

In our two fruit example,

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3.$$

A combination is a permutation with the number of ways to order the selected objects divided out. In our fruit example, there are P_2^3 permutations but 2! ways to arrange the chosen fruit, therefore the combination is $P_2^3/(2!) = 3$.

3.3 Working with Random Variables

A *random variable* links the outcomes of an experiment with numbers. When you roll a die, you look at it and read off a number. The random experiment is rolling a cube that has dots marked on its various faces. The outcome is actually the face that ends up on top. Without even thinking about it, we look at the face and read off a number. The number is the random variable. Random variables represent outcomes as numbers. The probabilities associated with outcomes are now transformed to probabilities associated with the corresponding random number. When rolling a die, we discuss the probability of getting a number from 1 to 6 rather than discussing the otherwise hard to describe outcomes. Similarly, if we select a card from a deck of cards, say the one with seven hearts on it, we recognize that as a 7 and the suit as hearts. Numbers that are determined by random mechanisms are called random variables.

In this chapter we use capital letters from the end of the alphabet to denote random variables and the corresponding lower case letters to denote their observed values. For example, the outcomes resulting from rolling a die can be represented by random variable X. X takes on any value from $1,2,3,4,5,6$. Once the die is rolled, we observe the outcome $X = x$, a specific value of X. If we observe a 4, we write $X = 4$. We say the random variable X takes on the value x with probability $1/6$ since all of the outcomes are equally likely. This is written $\Pr(X = x) = 1/6$ for $x = 1,...,6$. In subsequent chapters we will be more flexible in notation.

A random variable that has a finite or countable number of outcomes is called a *discrete random variable*. Since the outcomes can be listed, we can assign probabilities to each outcome. Random variables that yield random measurements typically do not have a countable number of outcomes. We can imagine the height of a person being 56.6πcm but we cannot count all such numbers. Random variables that can take on any value in some contiguous part of the real line are called *continuous random variables*. For continuous random variables, we cannot list all the outcomes so we must assign probabilities in some other way. In fact, we assign probabilities to intervals of numbers. Of course, it is fairly obvious that all random variables we can measure must be discrete random variables, because we are not capable of measuring heights like 56.6πcm. Nonetheless,

Table 3.2: *Probability mass function for X*

x	-2	0	3	4
$\Pr(X = x)$	0.25	0.1	.4	.25

when dealing with measurements, continuous random variables provide a valuable approximation to the extremely complicated discrete random variables that actually apply to measurements.

3.3.1 Probability Distributions, Expected Values

A probability distribution is a list of (numerical) outcomes and a means of specifying probabilities for the events containing the outcomes. It assigns probabilities to the outcomes of random variable X. *may need to explain this, not everyone may know*. From elementary physics, we can think of probability as a mass spread over the possible outcomes of a random variable X. X has associated with it a center of gravity. This is called the expected value of the random variable. For a random variable X with a discrete number of outcomes x this is

$$E(X) = \sum_{\text{all } x} x\Pr(X = x).$$

$\mathrm{E}(X)$ is a weighted average of all possible values of X. The weights (masses) are the probabilities.

EXAMPLE 3.3.1. A computer program is used by the Psycho Fiends Brigade to bill people who call their 1-900 number. The PFB's software is a complicated Bayesian belief network, but it has bugs in it. It charges 2 minutes less than a person called, the correct time, or either 3 or 4 minutes more than was called. The incorrect billings caused by the Psyco Fiends network has the probability distribution given in Table 3.2.

The expected value of X, written $E(X)$, is a measure of the center of the random distribution. It is computed as the weighted average of the outcomes x where the weights are given by the corresponding probabilities.

$$E(X) = \sum_{\text{all } x} x\Pr(X = x) = (-2)(0.25) + (0)(0.1) + (3)(0.4) + (4)(0.25) = 1.7.$$

On average, people are getting billed for 1.7 minutes more than they use. Thus, if the Psycho Fiends charged everyone 1.7 minutes less, they might make the District Attorney's office happy, although 65% of their clients would still be upset over being billed too much.

On occasion, we need to compute the expected value of a function of X. For example $E(X^2)$ is computed as

$$\begin{aligned} E(X^2) &= \sum_{\text{all } x} x^2\Pr(X = x) \\ &= (-2)^2(0.25) + (0)^2(0.1) + (3)^2(0.4) + (4)^2(0.25) \\ &= 8.6 \end{aligned}$$

In general,

$$E[g(X)] = \sum_{\text{all } x} g(x)\Pr(X = x).$$

Note that on the left hand side of the equation, we take the expected value of a function of the random variable X, say, $g(X)$. On the right hand side, for each specific outcome x of X, we use $g(x)$.

With regard to the Psycho Fiends call charges, the DA may care only about average charges, but

you as a customer care about being charged accurately. OK, you don't mind being undercharged, but you don't want to be overcharged, and the Fiends cannot stay in business if they undercharge everyone. Thus, we need to measure, not only the center of the distribution with $E(X)$, but also the variability associated with X. The *variance* of a random variable is a measure of spread (or *variability* or *dispersion*) of the distribution of the values of the random variable. The variance of a random variable X, is denoted $\text{Var}(X)$. Suppose X has mean, $E(X) = \mu$. Then the variance of X is defined as

$$\text{Var}(X) = E(X-\mu)^2 = \sum_{\text{all } x} (x-\mu)^2 \text{Pr}(X = x).$$

Note that since X is a random variable, $X - \mu$ is also a random variable, as is $(X-\mu)^2$, which is the squared distance between X and the center of its distribution, $E(X)$. The farther away X is from its mean μ, the larger the value of $(X-\mu)^2$. The expected value of $(X-\mu)^2$ measures the spread of the values of X about the mean μ, i.e., it is a measure of the average spread. This average is a weighted average with the weights being the probabilities that X takes on that specific value x. From physics, just as the expected value is the center of gravity of the probability mass function of X, the variance of X is the moment of inertia about the central axis (mean).

The $\text{Var}(X)$ for X defined in Example 3.3.1. is

$$\text{Var}(X) = (-2-1.7)^2(0.25) + (0-1.7)^2(0.1) + (3-1.7)^2(0.4) + (4-1.7)^2(0.25) = 5.71.$$

Note that while the units of X are minutes, this quantity has units of minutes^2.

The standard deviation of X is defined as the positive square root of the variance, and is denoted $s.d.(X)$. For this example, the standard deviation is

$$s.d.(X) = \sqrt{5.71} = 2.39.$$

Note that $s.d.(X)$ is a measure of variability that is again measured in minutes.

These definitions extend to functions of random variables. Suppose X is a random variable, a and b are constants, and we consider Y, a function of X, given by $Y = aX + b$. Y is a random variable. For example, X might be a random temperature measurement in degrees Celsius and $Y = 1.8X + 32$ is the same random temperature measured in degrees Fahrenheit. We find the expected value of Y as

$$E(Y) = E(aX + b).$$

It can be shown that

$$E(Y) = E(aX + b) = aE(X) + b$$

and that the variance and standard deviation are

$$\text{Var}(Y) = a^2 \text{Var}(X) \text{ and } s.d(Y) = |a| \, s.d.(X).$$

3.3.2 Independence, Covariance, Correlation

Suppose we record the time it takes an automobile to accelerate from 0 to 60 miles per hour, and also record its fuel consumption rate. Here we have defined 2 random variables, X is the time in seconds, and Y is the miles per gallon. A joint probability distribution of these random variables would give us all of the information about them. A joint probability distribution would specify all the outcomes for every pair (x, y) and specify the probabilities associated with such pairs. The joint distribution gives us relevant quantities of interest such as means and variances, as well as telling us whether events involving only X are independent of events involving only Y. For example, suppose we also record the age of the car. In later chapters, we will consider questions such as predicting how fast the car will achieve the 0 to 60 MPH speed if we know its age and fuel consumption rate, i.e. predict the value of one variable given the values of some other variable(s).

Table 3.3: *Joint probability mass function*

		X		
		-1	0	2
	1	0.1	0	0.3
Y	2	0	0.2	0
	3	0.2	0	0.2

EXAMPLE 3.3.2. *we need a story for this, why not the miles per gallon story.* Suppose X and Y have joint probability mass function given by Table 3.3. Here X is a measure of the number of seconds it takes a car to go from 0 to 60mph where 0 indicates 30 seconds, -1 is 20 seconds, and 2 is 50 seconds. Y is miles per gallon with $1 = 10$mpg, $2 = 20$mpg, etc.

We denote joint probabilities as $\Pr(X = x, Y = y) = p(x,y)$. So, $p(-1,2) = 0$, tells us that the probability $X = -1$ and $Y = 2$ is 0. If we wanted only information about X from this joint mass function, we would compute the marginal probability mass function (pmf) of X by summing out Y

$$\Pr(X = x) = \sum_{\text{all } y} p(x,y).$$

For example, $\Pr(X = -1) = 0.1 + 0.2 = 0.3$.

We can extend the definition of independence of events to include independence of random variables since random variables are based on events. Two random variables X and Y are independent if any events based on only X are independent of any events based on only Y. In other words, knowing anything about X does not change the probability for any event involving Y. In terms of probabilities, it is sufficient to have

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y).$$

In Example 3.3.2. , for $X = -1$ and $Y = 1$, we check independence

$$\begin{aligned} \Pr(X = -1, Y = 1) &= .1 \\ \Pr(X = -1) &= .3 \quad \text{and } \Pr(Y = 1) = 0.4 \end{aligned}$$

so

$$\Pr(X = -1)\Pr(Y = 1) = 1.2 \neq 0.1 = \Pr(X = -1, Y = 1).$$

Here are two events $(X = 1)$ and $(Y = 1)$ that are not independent, therefore X and Y are not independent.

We can also see whether X and Y have any sort of linear relationship. The *covariance* is a measure of linear relationship. Suppose X and Y are random variables with means μ_X and μ_Y respectively, the covariance between X and Y is given by

$$Cov(X,Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] = \sum_{(x,y)} (x - \mu_X)(y - \mu_Y)\Pr(X = x, Y = y).$$

Note that a special case of this is

$$Cov(X,X) = \text{Var}(X).$$

Another special case occurs when we have independence. If the random variables X and Y are independent, their covariance is 0. In general, the covariance can be positive or negative and its relative value depends on the units of X and Y. Positive covariances occur when large values of X occur with large values of Y and small values of X occur with small values of Y. Negative covariances occur when small values of X occur with large values of Y and large values of X occur with small values of Y. If X is the time in seconds to accelerate from 0 to 60 in our car example and Y is the

Table 3.4: *Joint probability mass function*

		X			
		-1	0	2	$\Pr(Y=y)$
	1	0.1	0	0.3	0.4
Y	2	0	0.2	0	0.2
	3	0.2	0	0.2	0.4
	$\Pr(X=x)$	0.3	0.2	0.5	1

fuel consumption rate in MPG, and we compute the covariance between X and Y, the value should be negative. Note that this covariance would change if we were to record Y in kilometers per litre.

To standardize the covariance measure, a unitless quantity called the *correlation* is defined,

$$Corr(X,Y) \equiv Cov(X,Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}.$$

The correlation is always a number between -1 and 1. The meaning of a correlation is the same as that of a covariance but in addition, if the correlation between X and Y is 1, we say there is a perfect linearly increasing relationship between X and Y. *will probably cause confusion with earlier use of X and Y* If the correlation is -1, this is a perfect linearly decreasing relationship. A perfect linear relationship between two random variables means that an increase of 1 unit in one random variable is associated with an exactly proportional change in the other. In our car example, this means that if we make a series of time observations on the car accelerating from 0 to 60 m.p.h and we record the fuel consumption rate in miles per gallon and kilometers per litre, the pairs of numbers should have a perfect linear relationship. We emphasize that the relationship is *linear* because the correlation does not identify perfect nonlinear relationships.

EXAMPLE 3.3.3. Let X and Y be random variables with joint probability mass function as defined in example 3.3.2, cf. Table 3.?. Then

$$\begin{aligned}
E(X) &= (-1)(0.3)+(0)(0.2)+(2)(0.5) = 0.7,\\
E(Y) &= (1)(0.4)+(2)(0.2)+(3)(0.4) = 2,\\
\text{Var}(X) &= (-1-0.7)^2(0.3)+(0-0.7)^2(0.2)+(2-0.7)^2(0.5) = 0.81,\\
\text{Var}(Y) &= (1-2)^2(0.4)+(2-2)^2(0.2)+(3-2)^2(0.4) = 0.8,
\end{aligned}$$

$$\begin{aligned}
&Cov(X,Y)\\
&= (-1-0.7)(1-2)(0.1)+(0-0.7)(1-2)(0)+(2-0.7)(1-2)(0.3)\\
&= (-1-0.7)(2-2)(0)+(0-0.7)(2-2)(0.2)+(2-0.7)(2-2)(0)\\
&= (-1-0.7)(2-2)(0.2)+(0-0.7)(3-2)(0)+(2-0.7)(3-2)(0.2)\\
&= -0.3,
\end{aligned}$$

$$\begin{aligned}
Corr(X,Y) &= \frac{-0.3}{\sqrt{(0.81)(0.8)}}\\
&= -0.37.
\end{aligned}$$

This correlation is quite a bit larger that -1, but indicates that small values of X tend to occur with large values of Y, as we can see from the probability distribution.

3.3.3 Expected values and variances for sample means

We now present some useful results for working with expected values, variances, and covariances of linear combinations of random variables. All of the results presented here generalize to more than 2

random variables. Suppose X_1, X_2, X_3, and X_4 be random variables and let a_1, a_2, a_3, and a_4 be real constants.

1. $E(a_1) = a_1$ and $E(a_1X_1) = a_1E(X_1)$.
2. $E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2)$.
3. $\text{Var}(a_1) = 0$ and $\text{Var}(a_1X_1) = a_1^2\text{Var}(X_1)$.
4. $\text{Var}(a_1X_1 + a_2X_2) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + 2a_1a_2Cov(X_1, X_2)$.
5. If X_1 and X_2 are independent, then $Cov(X,Y) = 0$. in which case 4 reduces to $\text{Var}(a_1X_1 + a_2X_2) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2)0$.
6. $Cov(a_1X_1 + a_2X_2, a_3X_3 + a_4X_4) = a_1a_3Cov(X_1, X_3)$
   $+ a_1a_4Cov(X_1, X_4) + a_2a_3Cov(X_2, X_3) + a_2a_4Cov(X_2, X_4)$.

The most basic data we collect is a random sample from a population. These are just independent random observations on the same population. Since the populations are the same, these observations will have the same distribution, which means that each of them would have the same mean (expected value) and variance.

We now concentrate on a function of random variables that is very important is statistical inference the sample mean. Let $X_1, X_2, ..., X_n$ be random variables each with the same population mean $E(X_i) = \mu$ for $i = 1, ..., n$. The average of these random variables is also a random variable. Define the *sample mean* as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We can find the expected value of this average of the random variables by finding the expected value of the linear combination

$$\begin{aligned} E(\bar{X}) &= E(\frac{1}{n}X_1 + \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n) \\ &= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) \\ &= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu \ \text{ (n terms)} \\ &= \frac{1}{n}n\mu \\ &= \mu \end{aligned}$$

When we have a random sample, we use $\bar{X}$ to estimate the population mean μ. The fact that $\text{E}(\bar{X}) = \mu$ indicates that $\bar{X}$ is a reasonable estimate of μ.

If we assume further that $X_1, X_2, ..., X_n$ are independent and have the same variance, say, $\text{Var}(X_i) = \sigma^2$ for $i = 1, ..., n$, we can compute the variance of $\bar{X}$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n) \\ &= \left(\frac{1}{n}\right)^2\text{Var}(X_1) + \left(\frac{1}{n}\right)^2\text{Var}(X_2) + \cdots + \left(\frac{1}{n}\right)^2\text{Var}(X_n) \\ &= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 \ \text{ (n terms)} \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

The variance of $\bar{X}$ is the variance of an individual observation, $\text{Var}(X_i)$ divided by n. When we

have a random sample, σ^2/n gives a measure of the variability of $\bar{X}$ as our estimate of μ. More commonly, we use the standard deviation $s.d.(\bar{X}) = \sigma/\sqrt{n}$. Regardless of the measure, $\bar{X}$ has much less variability than an individual observation X_i (depending on the size of n). Ultimately, after collecting data we compute $\bar{X}$ which means we we get to look at one observation on $\bar{X}$. Because $\bar{X}$ has less variability, that one observation on $\bar{X}$ is likely to be closer to μ than if we were to look at a single observation, say X_1.

Define the *sample variance* as

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 + -\bar{X})^2 \cdots + (X_n - \bar{X})^2}{n-1}.$$

The rules for expected values can be used to establish that

$$\mathrm{E}(S^2) = \sigma^2,$$

although doing so is more laborious than our results for the sample mean. The degrees of freedom (df) for a variance estimate is an index of how much information is in the estimate. For the sample variance the degrees of freedom are $n-1$ because there are n observations but we lose one degree of freedom for not knowing $\mathrm{E}(X_i) = \mu$ and replacing it with its estimate $\bar{X}$. In Chapters 4 and 9 we will average several independent variance estimates so their degrees of freedom will be the sum of the df for the individual estimates.

3.4 Binomial Distribution

Certain distributions arise repeatedly in statistical applications. These distributions are given names. The binomial distribution is one that arises when we have a fixed number of independent and identical random trials that result in one of two types of outcomes. (Binomial comes from roots meaning two names.) We are interested in the random numbers of each type of outcome. Note that with a fixed number of trials, if we know the number of times one outcome occurs, we can immediately figure out the number of the other outcome. Examples are

testing a batch of identical components and counting the number of components that pass (or fail) the test,

checking airline scheduling accuracy and counting the number of flights that left on time.

checking whether a community meets the standards for group immunity against measles, and counting the number of people who have been vaccinated against measles

Each of these situations has only two possible outcomes. Generally, these are termed successes or failures. To understand the distribution of the number of successes in these situations, consider the first example. Suppose we have a batch of 4 resistors and we are going to test each one to see whether or not it works. Each resistor is a trial. The probability that a resistor is good is p and the probability that a resistor is bad is $1-p$. We usually call the outcome we are interested in a "success". In this case, $\Pr(success) = p$ and the probability of the complementary event, $\Pr(failure) = 1-p$. Let X count the number of good resistors, so X can take on the values $0,1,2,3,4$. Define events $G =$ resistor is good and $D =$ resistor is defective. The event $\{X = 0\}$ means that none of the resistors are good. Using the independence of the resistors, we find the probability $X = 0$ as

$$\begin{aligned} \Pr(X=0) &= \Pr(DDDD) \\ &= \Pr(D)\Pr(D)\Pr(D)\Pr(D) \\ &= (1-p)(1-p)(1-p)(1-p) = (1-p)^4. \end{aligned}$$

The event $\{X = 1\}$ means that one of the resistors is good and the other 3 are defective. This can happen in many ways. We can think of this as the number of ways to select one good resistor

Table 3.5: *Probability mass function for the number of good resistors*

X	$\Pr(X=x)$
0	$\binom{4}{0}p^0(1-p)^4$
1	$\binom{4}{1}p^1(1-p)^3$
2	$\binom{4}{2}p^2(1-p)^2$
3	$\binom{4}{3}p^3(1-p)^1$
4	$\binom{4}{4}p^4(1-p)^0$

from the resistors. Using the addition rule and independence, $\Pr(X=1)$ is

$$\begin{aligned}
&\Pr(X=1)\\
&\quad = \quad P\{\ GDDD \text{ or } DGDD \text{ or } DDGD \text{ or } DDDG\ \}\\
&\quad = \quad \Pr(GDDD)+\Pr(DGDD)+\Pr(DDGD)+\Pr(DDDG)\\
&\quad = \quad p(1-p)^3+(1-p)p(1-p)^2+(1-p)^2p(1-p)+(1-p)^3p\\
&\quad = \quad 4p(1-p)^3\\
&\quad = \quad \binom{4}{1}p(1-p)^3.
\end{aligned}$$

The $\Pr(X=2)$ is the probability of having 2 good resistors.

$$\begin{aligned}
&\Pr(X=2)\\
&\quad = \quad P\{GGDD \text{ or } GDGD \text{ or } DGGD \text{ or } DGDG \text{ or } DGGD \text{ or } DDGG\}\\
&\quad = \quad \Pr(GGDD)+\Pr(GDGD)+\Pr(DGGD)\\
&\qquad\quad +\Pr(DGDG)+\Pr(DGGD)+\Pr(DDGG)\\
&\quad = \quad 6p^2(1-p)^2\\
&\quad = \quad \binom{4}{2}p^2(1-p)^2,
\end{aligned}$$

where the last line is given as an alternative to get the probability. Rather than listing each outcome, we can think of the number of ways of having 2 defective resistors in 4 resistors, $\binom{4}{2}=6$. The probability of any one of these outcomes with 2 good and 2 bad resistors is $p^2(1-p)^2$ and there are 6 ways that this can happen. We compute the remaining probabilities in the same manner and they are listed in Table 3.5.

In general, we do not want to list every single outcome every single time we have a situation like this, and especially not if the number of trials n is large! The general functional form for this type of probability distribution, called the *binomial* distribution is

$$\Pr(X=x)=\binom{n}{x}p^x(1-p)^{n-x},\ \text{for } x=0,1,...,n,\ 0\le p\le 1.$$

We use the notation $X\sim Bin(n,p)$ to mean that random variable X is binomially distributed with parameters n and p.

EXAMPLE 3.4.4. When one of us was a student, she owned a car that was designed such that every time it rained, water leaked through and took out a fuse or 2 or 3. Fortunately, she lived in the desert, but in any case, she was forced to buy fuses in boxes of 10. On any given box, there was a warranty that stated that she could get her money back if 2 or more fuses were defective. The probability that a fuse is defective is .01. What is the probability that she would get my money back?

First check to see whether this fits a binomial situation. There are $n=10$ fuses. Each fuse is a

trial. We can reasonably assume for the moment that the fuses are defective independently of one another. Each fuse is defective with probability $p = 0.01$. Let X count the number of defective fuses in a box, then $X \sim Bin(10, .01)$. The event $\{X \geq 2\}$ is the event that the money-back warranty would have to be honored. So we find $\Pr(X \geq 2)$

$$\Pr(X \geq 2) = \Pr(X = 2) + \Pr(X = 3) + \Pr(X = 4) + \cdots + \Pr(X = 10).$$

It is easier to look at the complementary event

$$\begin{aligned} \Pr(X \geq 2) &= 1 - \Pr(X \leq 1) \\ &= 1 - [\Pr(X = 0) + \Pr(X = 1)] \\ &= 1 - \left[\binom{10}{0} (0.01)^0 (1-0.01)^1 0 + \binom{10}{1} (0.01)^1 (1-0.01)^9 \right] \\ &= 0.004 \end{aligned}$$

The mean and variance of $X \sim \text{Bin}(n, p)$ are difficult to compute from the definitions in Subsection 3.3.1, but are quite simple to compute using results from Subsection 3.3.3. We merely state the results:

$$\text{E}(X) = np \qquad \text{Var}(X) = np(1-p).$$

Note also from the definition of a binomial, that if $X_1 \sim \text{Bin}(n_1, p)$ and independently $X_2 \sim \text{Bin}(n_2, p)$ then $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$.

mean and variance of a binomial. sums of binomials.

3.5 Poisson distribution

A probability distribution that is closely related to the Binomial is called the Poisson distribution. The Poisson distribution is used to model the occurrence of very rare events, e.g., rare diseases and the number of defects in high quality products. While the proportion of defects may be very small, when there are a very large number of trials the number of defects that occur can be substantial. An advantage of the Poisson distribution is that it can be used to model product defects when we have no clear idea of the number of units that have been produced that could be defective. For example, one might consider the number of painting defects on a piece of sheet metal. How many locations are there for the painting to be defective? Innumerable. The probability of having a painting defect at any one point is very small but the number of defects can still be substantial.

Suppose we have a binomial with a very large number of trials n and a very small probability of occurrence p. Since n is large and p is small, their product, which is the expected number of defectives, can be of moderate size, say $\mu = np$. The binomial variance is $np(1-p) = \mu(1-p)$ but since p is very small, $1-p$ is essentially 1 and the variance is essentially μ. In other words, if Y has a Poisson distribution with parameter μ, write $Y \sim P(\mu)$. The expected value is $E(Y) = \mu$ and $Var(X) = \mu$.

3.6 Normal Distribution

Often, our random variables of interest are continuous. Ideally, a continuous random variable can take on any value in an interval. The actual value of such a variable is limited by the precision of the measuring instrument. Examples are time, temperature, weight, and strength. All are measured to the nearest unit. When we say that the temperature is $70°F$, we mean that it is somewhere in an interval about $70°F$, say between $69.5°$ and $70.5°$. There are many continuous distributions that are useful for modeling continuous data. The most common of these is the normal distribution.

The normal distribution arises naturally in many physical experiments. In particular, the normal distribution can provide a close approximation to both the Binomial and Poisonn distributions when their expected values are reasonably large.

Discuss standard normal and need for more generality. Examples and pictures, Maybe use the roller coaster riders.

Chapter 4

Control Charts

In this chapter, we examine *control charts*. Control charts are used to establish whether a process is in statistical control. Theoretically, a process is in statistical control when it generates a sequence of random observations that are independent with the same statistical distribution for each observation (iid). Unfortunately, it is impossible to tell whether a process is iid by observing it. Control charts provide an *operational definition* (one that can be checked against observations) for whether observations are close enough to iid for us to treat them as such. Since we are not proving that the observations are iid, we define the process to be *under control* when the operational definition is satisfied. Here we discuss components for a definition of "under control" that have been found useful historically, along with their statistical motivations. Essentially, we use observable properties of iid sequences to define what it means to be under control.

Control charts are merely plots of the data or data summaries with control limits superimposed on the plots. The control limits are used to identify processes that are out of statistical control. **Control charts were developed by the founder of statistical quality control, Walter A. Shewhart.** His ideas were systematically introduced in Shewhart (1931). *The ultimate goal of establishing that a process is under statistical control is that it provides a basis for making accurate predictions about the future product from the process.*

EXAMPLE 4.0.1 Table 4.1 contains 50 observations on a person's diastolic blood pressure. This is actually a subset of a collection of 460 blood pressures taken on this individual twice each day as reported by Shumway (1988). The individual in question determines a process for producing diastolic blood pressures. We will investigate these 50 observations to examine whether the process is in statistical control. We will also consider control charts for blood pressures that were measured four times on a regular basis but were not measured every day. Diastolic blood pressures of 94 or greater are considered high; we will also consider control charts that examine the frequency of high blood pressure values. □

Control limits are based on a simple idea. Let y be a random data point, let the population mean of such observations be $\mathrm{E}(y) = \mu$, and let the population variance be $\mathrm{Var}(y) = \sigma^2$ so that the population standard deviation is σ. Control limits are based on the fact that y will very probably fall between $\mu - 3\sigma$ and $\mu + 3\sigma$. If y happens to follow a normal distribution, the probability of being in this interval is 99.7%. Chebeshev's inequality ensures that the probability is always at least

Table 4.1: *Diastolic Blood Pressures*

	0	1	2	3	4	5	6	7	8	9
00	105	92	98	98	98	95	92	94	92	92
10	95	92	94	93	98	96	92	94	98	91
20	92	90	88	97	93	95	96	84	93	92
30	94	90	90	85	94	90	95	94	95	85
40	94	92	94	92	92	90	94	90	85	90
Read across and then down.										

$1-(1/3^2) \doteq 88.9\%$ and, under weak conditions discussed in Pukelsheim (1994) or briefly in Christensen (1996, Section 1.3), this can be raised to $1-[(2.25)3^2]^{-1} \doteq 95.1\%$. When an observation does not fall within the interval, something unusual, i.e., a *special cause* has occurred. (*Common causes* are those that determine the inherent variability of the process and can only be controlled by management.) Control charts differ depending on the exact nature of the datum y. Often, the datum is actually the average of several observations. Sometimes the datum is the number of defectives among a fixed number of items. Sometimes the datum is the number of defectives in a large but unspecified number of items. Moreover, μ and σ are, in practice, unknown and must be estimated. Various control charts use different estimates of μ and σ.

It is not our goal to give an exhaustive discussion of control charts but rather to present examples that introduce some of the main ideas and types of control charts in a simple manner. To this end, we use notation that is consistent with the rest of the book but therefore is less consistent with traditional discussions. We also present charts that are simple but less commonly used. Rather than illustrating all of the more commonly used charts, we often merely discuss their differences from those presented here. For a detailed discussion of control charts see, for example, Ryan (1989).

Control charts and the corresponding control limits have clear connections to statistical prediction intervals and statistical tests. The charts are computed using assumptions that, ideally, should be subjected to verification. We could clearly do more sophisticated analyses than control charting but at the danger of intimidating and even driving away users of statistics. Control charts should perhaps be viewed as simply a plot of the data along with an estimate of the average level of the process and a *rough* measure of how much variability one reasonably expects to see. In practice, control charts are often used by people with a minimum of statistical training. They are rough and ready (quick and dirty?) procedures. In practice, the choice is often between a simple data analysis and no data analysis. Moreover, one can, and probably should, take a different philosophical approach to control charts that calls in question the association with statistical tests and prediction intervals. We return to the philosophical issue later.

Basically the operational definition of whether a processes is under control is that it stays within the control limits of

- a means chart and a dispersion (s, R, or S^2) chart for rational subgroup data or
- an individual chart and moving range chart for individual observations or
- a p, Np, c or u chart for attribute data.

To that operational definition one can add extra conditions like those described in Nelson (1984) or possibly versions of EWMA charts or CUSUM charts. All of these procedures are discussed in this chapter.

Perhaps more important than the details of constructing control charts is getting people to graph their data and look at it.

4.1 Individuals Chart

EXAMPLE 4.1.1 *An I (individuals) chart*. Consider the single sample of data presented in Table 4.1. From these data compute $n = 50$, $\bar{y}_\cdot = 92.880$, and $s = 3.799$. We want control limits of $\mu \pm 3\sigma$. These are estimated with $\bar{y}_\cdot \pm 3s$ so the estimated upper control limit is

$$UCL = 92.880 + 3(3.799) = 104.277$$

and the estimated lower control limit is

$$LCL = 92.880 - 3(3.799) = 81.483.$$

The control chart is displayed in Figure 4.1. The plot displays all of the observations, the upper and lower control limits are given as horizontal lines, and a horizontal line is used to indicate the estimated mean value of 92.88.

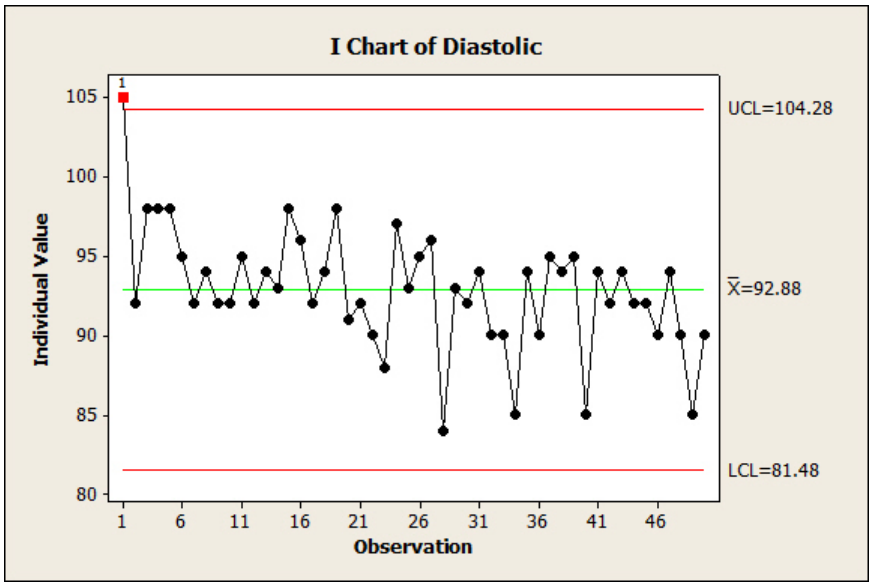

Figure 4.1: *I (Individuals) Chart for Table 4.1.*

We will defer a detailed discussion of interpreting control charts until the next example but note that the very first observation is beyond (apparently on) the upper control limit and that four of the first five observations are at a substantial distance above the mean line. Both of these facts indicate that the process is not in control. To get the process under control, some action is needed. Perhaps the initial high readings are due to the patient going on a beer, licorice, and salted peanut binge just before the study started. This is one possible explanation for the process being out of control. Other possible explanations must be sought out.

Traditionally, the sample standard deviation s is not used for estimating σ in this control chart. The traditional method involves going through the data and averaging the distances between adjacent data points (*moving ranges*). (Recall that distances are always positive.) This average distance is then divided by a tabled adjustment factor to give an estimate of σ. Alternatively, the sample standard deviation s is sometimes divided by a different tabled adjustment factor to give an unbiased estimate of σ. In deriving all of the estimates of σ, the observations are assumed to be independent. In addition, the tabled adjustment factors both depend on the data being normally distributed. The assumption of independence is frequently a bad one for data used in control charts and we will see later that it is a particularly bad assumption for the diastolic blood pressure data.

Figure 4.2 gives a control chart based on the moving ranges. It is a plot of the distances between consecutive observations and is designed to evaluate whether the variability of the process remains constant. The average of the moving ranges can be used to develop an alternative estimate of the standard deviation that is not as liable to bias due to a shift in the process mean but it requires an assumption of normal data. The center line is the mean of the moving ranges and the standard deviation of a moving range is a multiple of σ and so is estimated using a tabled multiplier of the mean moving range. (This can all be related to the theory behind Tukey's HSD multiple comparison method.) *Minitab commands for control charts are given in Section 4.8.* □

Alternatively, the variance can be estimated by the squared distance between consecutive points (squared moving ranges),

$$\tilde{\sigma}^2 \equiv \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2.$$

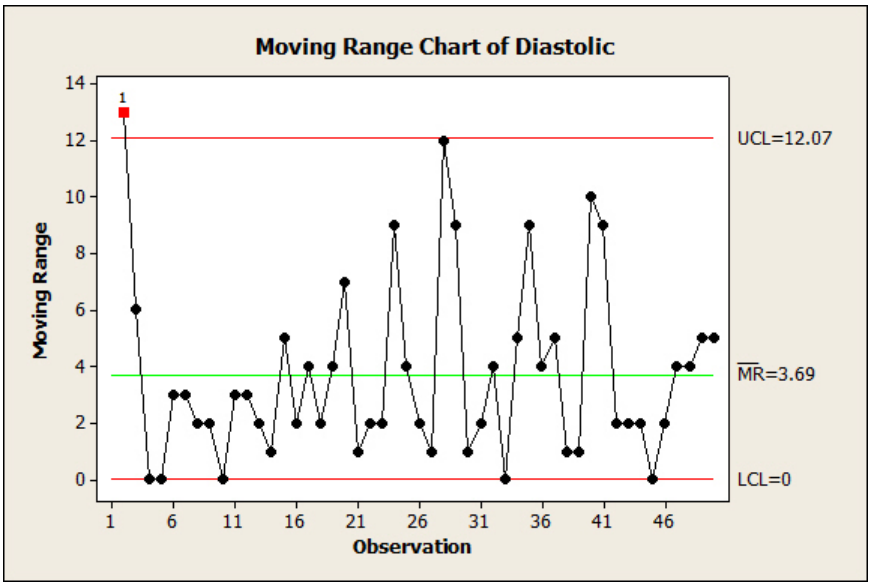

Figure 4.2: *Moving Range Chart for Table 4.1.*

This can be shown to have

$$\mathrm{E}\left(\tilde{\sigma}^2\right) = \sigma^2.$$

If the process is subject to a mean shift at some time, as a function of the moving ranges, $\tilde{\sigma}^2$ is subject to far less bias due to the mean change than is s^2.

The key point from Example 4.1.1 is that when the process is out of control, action must be taken to bring the process under control before attempting to improve the process by reducing variability. As with any control chart, *once the process is brought under control, the control chart needs to be recomputed*.

4.2 Means Charts and Dispersion Charts

Perhaps the most commonly used control charts are based on data collected in *rational subgroups*. An example of such data is given in Table 4.2. This contains 19 groups of four blood pressure readings. The four readings were taken under *essentially identical conditions*. Two charts are traditionally constructed: one that plots the means for the various samples and one that plots the ranges of the various samples. Recall that the range of a sample is simply the largest observation in the sample minus the smallest observation in the sample. The *means chart* is a tool to identify changes in the process mean but it can also show changes in process variability. In Subsection 9.1.2 we will discuss the ability of the means chart to detect lack of independence. The *range chart* is designed to examine the process variability. Alternatives to the range chart are the *sample standard deviation chart* and the *sample variance chart*. In the next example, we examine the means chart, the sample standard deviation chart, and the range chart. The last two of these are typically referred to as s and R charts, respectively. The means chart is traditionally called an $\overline{X}$ (X-bar) chart but we always use y for the variable being analyzed, so we would refer to this as a $\bar{y}$ chart. (Software may label them as $\overline{X}$ or X-bar.)

The standard advice is that *one should never interpret a means chart unless a corresponding dispersion chart (range, standard deviation, or variance) has already established that the variability is under control*. While means charts are sensitive to a lack of control in the variability of the process,

Table 4.2: *Nineteen Samples of Diastolic Blood Pressures*

Group	Data					Mean	Std. Dev.	Range
i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	N	$\bar{y}_{i\cdot}$	s_i	R_i
1	105	92	98	98	4	98.25	5.32	13
2	95	96	84	93	4	92.00	5.48	12
3	90	97	97	90	4	93.50	4.04	7
4	90	88	86	90	4	88.50	1.915	4
5	100	92	92	90	4	93.50	4.43	10
6	78	82	82	84	4	81.50	2.52	6
7	90	83	83	87	4	85.75	3.40	7
8	95	85	88	90	4	89.50	4.20	10
9	92	90	87	85	4	88.50	3.11	7
10	84	85	83	92	4	86.00	4.08	9
11	90	87	84	90	4	87.75	2.87	6
12	93	88	90	88	4	89.75	2.36	5
13	92	94	87	88	4	90.25	3.30	7
14	86	87	91	88	4	88.00	2.16	5
15	90	94	94	82	4	90.00	5.66	12
16	95	95	87	90	4	91.75	3.95	8
17	96	88	90	84	4	89.50	5.00	12
18	91	90	90	88	4	89.75	1.258	3
19	84	82	90	86	4	85.50	3.42	8
Total					76	89.434	4.848	7.947

to know whether a problem identified in a means chart is a problem with the level of the process or a problem with variability, one needs to examine both charts.

EXAMPLE 4.2.1 The data in Example 4.1.1 are actually a subset of 460 blood pressure measurements. The complete data were sampled to obtain the data given in Table 4.2. Frequently it is impractical or wasteful to take measurements on every item produced. Going to a doctor's office twice every day to have your blood pressure measured is certainly very inconvenient and is likely to be quite expensive. (Much nicer to buy your own wrist cuff.) To simulate realistic sampling from a process, the 460 observations were divided into 18 consecutive groups of 25 with a final group of 10. The first four observations from each group were taken as the sample from the group. The data are given in Table 4.2 along with the group mean, standard deviation, and range.

This method of obtaining data is similar in spirit to evaluating an industrial process in which a worker produces 25 units per day and measuring the first four units produced each day. Alternative sampling schemes might be to sample units produced at four specified times during the day or to take a random sample of four units from the 25 produced in a day. If one samples at specified times of day, the times may need to be worked into the analysis, in which case the data structure would differ from Table 4.2. In any case, *the groups must be formed in some rational manner*. Shewhart emphasized that they should be taken under *essentially identical conditions*. This leaves room for detecting effects that appear when conditions are not identical. (Sampling at specified times of day to form a group is a bad idea if the process can change in the course of a day.) Shewhart also suggested that one should probably have 25 or more groups (more than our 19 groups) before declaring a process to be under control, i.e, we can potentially find our process to be out of control but no matter what the data tell us we should be hesitant to say that it is under control.

The last row of Table 4.2 includes the sample standard deviation of the 76 observations computed by ignoring the group structure, 4.848, and the grand mean of all observations, $\bar{y}_{\cdot\cdot} = 89.434$, which is used to estimate μ. It also contains the mean of the sample ranges, $\overline{R} = 7.947$. The *pooled estimate of the variance* σ^2 is just the average of the s_i^2s. This is often called the *mean squared error (MSE)*. The value is 14.48 and the pooled estimate of σ is

$$\sqrt{MSE} = 3.81.$$

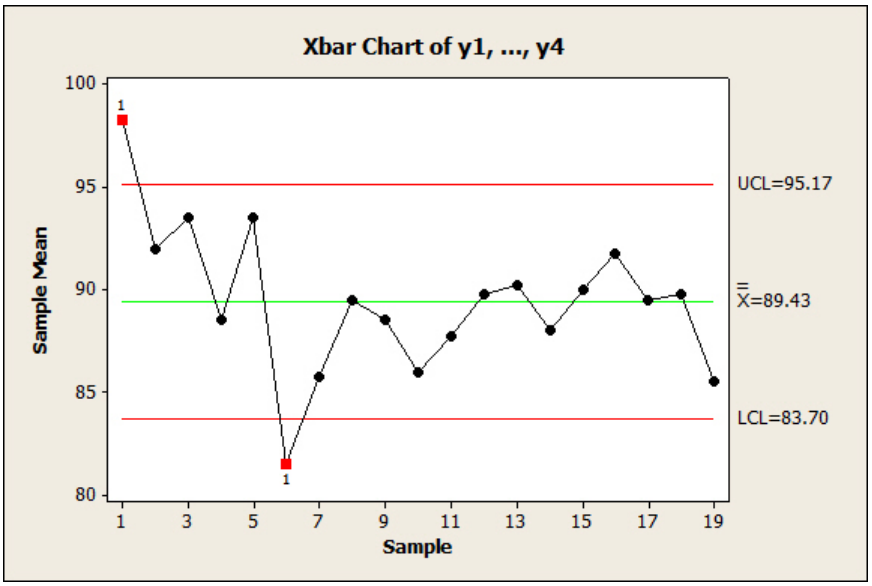

Figure 4.3 *Means Chart for the* $(i, \bar{y}_{i\cdot})$ *data of Table 4.2 using* $\sqrt{MSE}$. (The label at the top of the chart is created by Minitab and refers to the four columns of numbers in the table that are y_{i1} through y_{i4}.)

Note that for an iid process, if we select groups of observations from the process using any selection criteria that does not depend on the actual values of the observations, then each datum is an independent observation from the process (population).

The control chart for the sample means is given in Figure 4.3. For a process under control, a typical observation y_{ij} is assumed to have mean $\mathrm{E}(y_{ij}) = \mu$ and $\mathrm{Var}(y_{ij}) = \sigma^2$. The observations plotted in the control chart are the group mean values $\bar{y}_{i\cdot}$. The mean values have $\mathrm{E}(\bar{y}_{i\cdot}) = \mu$ but $\mathrm{Var}(\bar{y}_{i\cdot}) = \sigma^2/4$ because there are 4 observations in each mean. The upper and lower control limits are determined by the standard deviation of $\bar{y}_{i\cdot}$ which is $\sigma/2$. The control limits are $\mu \pm 3(\sigma/2)$. Using the estimated parameters from the previous paragraph, the estimated control limits become

$$UCL = \bar{y}_{\cdot\cdot} + 3\sqrt{\frac{MSE}{4}} = 89.434 + 3\frac{3.81}{2} = 95.15$$

and

$$LCL = \bar{y}_{\cdot\cdot} - 3\sqrt{\frac{MSE}{4}} = 89.434 - 3\frac{3.81}{2} = 83.72.$$

In general, if each sample has size N, the estimated control limits are

$$\bar{y}_{\cdot\cdot} \pm 3\sqrt{\frac{MSE}{N}}.$$

Using the MSE as an estimate of σ^2 assumes that the observations within each group are uncorrelated (or independent). This is often a poor assumption; we will return to it later. (Independence between groups of observations is not crucial but, oddly, because groups tend to be more widely separated in time or other factors, the assumption of independence between groups is frequently reasonable.)

The control chart in Figure 4.3 is nonstandard. More commonly, estimates of σ other than $\sqrt{MSE}$ are used. The alternative estimates involve using tabled adjustments. The tabled adjustments assume that the observations within each group are both independent and have normal distributions. Often, σ is estimated by the average of the sample ranges, $\bar{R} = 7.947$, divided by a tabled constant.

This estimate is unbiased because, for independent normal observations, the expected value of the range of a sample is σ times this constant that depends only on the sample size. Alternatively, σ is sometimes estimated by $\sqrt{MSE}$ divided by a tabled constant. Although MSE is an unbiased estimate of σ^2, $\sqrt{MSE}$ is not an unbiased estimate of σ. The tabled constant corrects for the bias that occurs when the observations are independent and normal. (Even though a **little** *statistical* bias is not really such a bad thing in an estimate.) The means chart using the bias corrected estimate of σ based on $\sqrt{MSE}$ is very similar to Figure 4.3 because, with a large number of degrees of freedom for error, the bias correction factor is very close to one. (The degrees of freedom for MSE is the sum of the df for each s_i^2, hence is $n(N-1)$.)

Traditionally, means charts use $\overline{R}$ in estimating σ and R charts are used to evaluate dispersion. Interestingly, *Shewhart (1931) never uses such charts*. I presume their use was popularized because ranges are easier to calculate by hand than standard deviations but also because, for small group sizes like $N = 4$, the inefficiency introduced by using the range is small. (Although Mahmoud et al., 2010, take strong exception to this statement.) Ranges should not be used unless group sizes are very small and, with the accessibility of electronic calculation, *there seems to be little point in ever using ranges.* (Moving ranges for individuals are another story.)

As discussed earlier, we have to know the distribution of the y_is to know the probability that an individual observation, one that is under control, will appear to be under control. If we are plotting observations with normal distributions, the probability is 99.77%. Chebeshev's inequality assures us that the probability is at least 88.9% and, with additional assumptions, at least 95%. One advantage of means charts is that means tend to have distributions more like the normal than do individual (unaveraged) observations. Thus we can be more confident of having a relatively large probability of identifying means as being under control when they truly are under control.

Return now to interpreting Figure 4.3. Two of the observations, the first and the sixth, are beyond the control limits. The first three points and cases 3, 4, and 5 both have the property that two of the three consecutive points are more than two standard deviations above the center line of 89.43. (Two standard deviations above the center line is $89.43 + (2/3)(95.15 - 89.43) = 93.24$). Each of these events is highly unlikely when the group means are independent and normally distributed. Moreover, four of the first five points are more than one standard deviation above the center line. Again, this event is highly unlikely if the group means are independent and normally distributed.

As mentioned earlier, we should not interpret the means chart unless a corresponding dispersion chart establishes that the variability is under control. We now examine dispersion charts for these data. We will see that the dispersion seems to be controlled.

Figure 4.4 is a control chart of the sample standard deviations. It is used to detect changes in the variation of the process. The points plotted are just the s_is from Table 4.2. The center line at 3.521 is just $\sqrt{MSE}$ times the ratio of two bias adjustment factors. $\sqrt{MSE}$ is divided by one adjustment factor to get an unbiased estimate of σ and then multiplied by another adjustment factor so that the center line has the same bias as the s_is that are being plotted. In this way, the center line is an estimate of the expected value of the s_is. The control limits are the expected value of s_i plus and minus three times the standard deviation of s_i. For independent and normally distributed data, the standard deviation of s_i is a tabled multiple of σ that depends on the group sample size. The estimated control limits are 3.521 plus and minus three times the tabled value times the estimate of σ. Sometimes, as in this chart, the computed lower control limit is a negative number. Standard deviations cannot be negative so in such cases the lower control limit is taken to be zero. The s chart shows no lack of control.

The R chart in Figure 4.5 is an alternative control chart for detecting changes in the variation of the process. In this chart the ranges of the samples, as reported in Table 4.2, are used to indicate variability. The center line of the chart is the average of the sample ranges from Table 4.2, $\overline{R} = 7.947$. The control limits are the expected value of a range plus and minus three times the standard deviation of a range. For independent and normally distributed data, the standard deviation of a range is a tabled multiple of σ that depends on the group sample size. The estimated control limits in Figure 4.5 are 7.947 plus and minus three times the tabled value times the estimate of σ. In

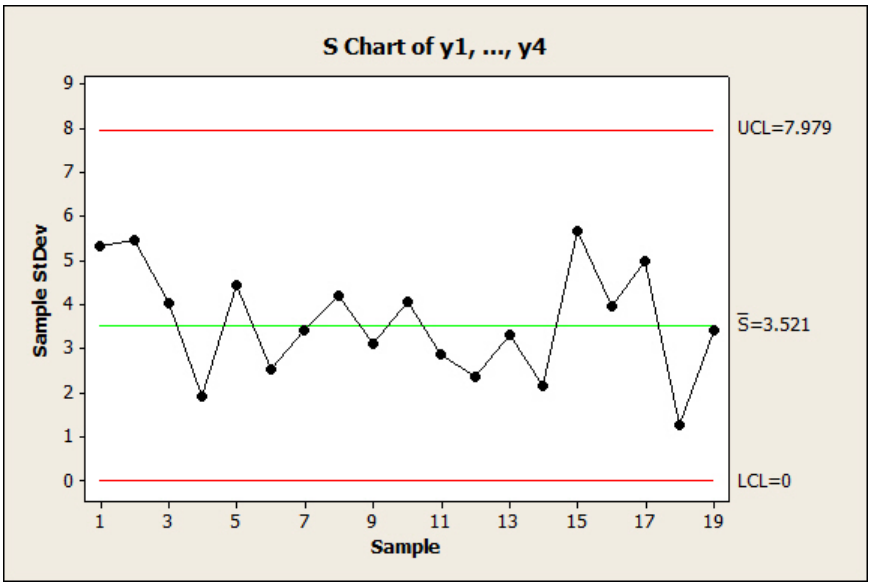

Figure 4.4: *s Chart for the* (i, s_i) *data of Table 4.2.*

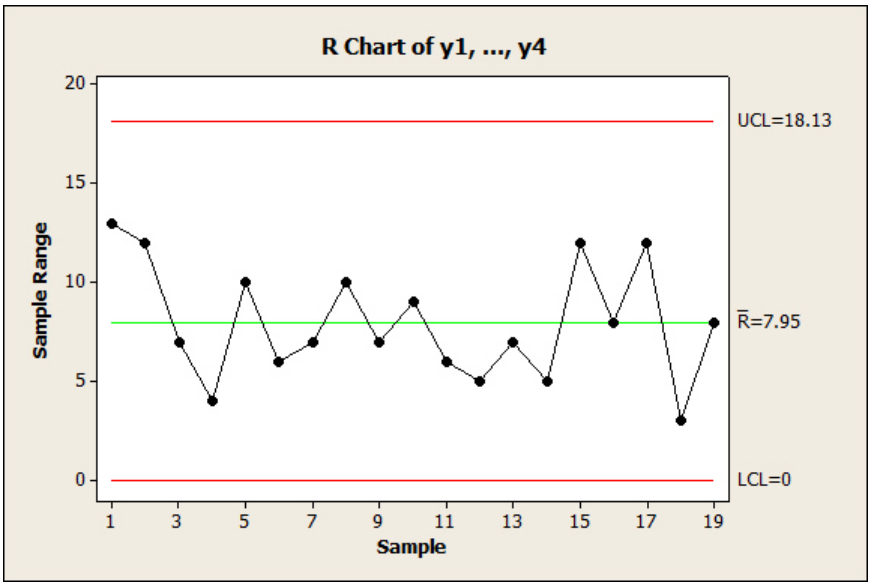

Figure 4.5: *R Chart for the* (i, R_i) *data of Table 4.2.*

this chart, the estimate of σ was taken as the appropriate multiple of $\overline{R}$. As with the *s* chart, if the lower control limit is negative, it is adjusted to zero because a range cannot be negative. The *R* chart displays no lack of control. □

Although they do not seem to be as popular as *s* and *R* charts, mathematically it is easy to create a *variance* (s^2) *chart*. If the variance in each group is the same,

$$\sigma^2 = \mathrm{E}(MSE) = \mathrm{E}(s_i^2),$$

so the chart plots the pairs (i, s_i^2) with center line *MSE*. With independent normal observations,

$(N-1)s_i^2/\sigma^2$ has something known as a chi-squared distribution with $N-1$ degrees of freedom, written

$$(N-1)s_i^2/\sigma^2 \sim \chi^2(N-1).$$

From properties of the $\chi^2(N-1)$ distribution we get

$$(N-1)\mathrm{E}[s_i^2]/\sigma^2 = N-1 \qquad \text{and} \qquad (N-1)^2\mathrm{Var}[s_i^2]/\sigma^4 = 2(N-1).$$

From this it follows that

$$\mathrm{E}[s_i^2] = \sigma^2 \qquad \text{and} \qquad \mathrm{Var}[s_i^2] = 2\sigma^4/(N-1).$$

(The first of these holds without the normality assumption that leads to the chi-squared distribution.) Estimating σ^2 with MSE we get control limits of $MSE \pm 3MSE\sqrt{2/(N-1)}$ and no need for finding correction factors.

There are several criteria (tests) commonly used with control charts to evaluate (determine) whether a process is under control.

1. The first test is obviously whether any points fall beyond the control limits. (Shewhart refers to this as Condition I.)
2. A second indication is the existence of 9 or more consecutive points on the same side of the center line.
3. A third indication is a trend of 6 or more points that are all increasing or decreasing. (Some authors require 7 points for a trend.)
4. A periodic, up and down cycling in the plot indicates nonrandomness and lack of control. Having 14 consecutive points that alternate up and down, i.e., no three points are increasing or decreasing, indicates a problem.
5. Too many, say 15, consecutive points within one standard deviation of the center line is a problem.
6. Having 8 consecutive points beyond one standard deviation of the center line is another problem.
7. Having two of three consecutive points beyond two standard deviations is a problem.
8. Having four of five consecutive points on the same side of the center line and beyond one standard deviation is a problem.

For more details, see Nelson (1984).[1]

In constructing means charts, Shewhart (1931) emphasizes that the observations must be divided into *rational subgroups* that have data generated under *essentially identical conditions*. Rational subgroups must be determined by the subject matter. If the observations are taken in time sequence, it is rational to group observations made at similar times. If 100 observations are the result of having four observations on each of 25 workers, it is rational to group the observations by worker. Shewhart (1931) also emphasizes the use of small subgroups because they are sensitive to fleeting changes in the process. $N = 4$ observations per group is very commonly used. Shewhart (1939, p. 37) recommends that one have *at least* 25 groups of 4 observations that appear to be in control *before drawing the conclusion that a process is under control*.

Control charts are used for two purposes: *process analysis* and *process control*. In our examples we discuss process analysis. We have past data on a process and seek to determine whether the process was in control when the data were obtained. Process control also uses past data to form a control chart but the center line and control limits are extended into the future so that new data can be plotted as they arrive. As long as the new data are consistent with a process in control, the process

[1] Stationary processes as discussed in Chapter 6 are "under control" but could easily violate these conditions.

is allowed to continue. When the new data indicate that the process is out of control, the process is adjusted to put it back in control. When the process is adjusted, control charts must be recomputed.

Once the process is under control, the goal should be to reduce variability while keeping the process on target, i.e., focused on the ideal specification. Staying on target is usually easier to achieve than reducing variability.

The data used for constructing a means chart has the same structure as data used in one-way analysis of variance, cf. Section 9.1. Also, Ott's *Analysis of Means* procedure, cf. Christensen (1996, Subsection 6.5.1), provides a test for equality of means that is directly related to means charts.

4.2.1 Process Capability

In many applications, notably industrial production, a process is required to produce items that meet certain specifications. Typically, these specifications come in the form of specification limits that items must fall between. It becomes of considerable importance to recognize whether a process is capable of meeting such specifications. Processes that are not under control have no capability to meet any specifications.[2] The capability of a process that is under control is defined to be $\mu \pm 3\sigma$. The process can be expected to produce items within these limits. If the observations are independent and normally distributed, 99.7% of the items produced will fall within the $\mu \pm 3\sigma$ interval. *If the interval defined by the specification limits contains the process capability interval, the process is capable of meeting specifications.* In other words, if the upper specification limit is above $\mu + 3\sigma$ and the lower specification limit is below $\mu - 3\sigma$, the existing process will meet the specifications (at least 99.7% of the time for normal, independent data). If the specification interval partially overlaps or misses the process capability interval, the process needs to be changed before it can meet specifications. *Until the process is brought under control with the capability of meeting specifications, all items produced must be inspected individually to see whether they meet specifications. When the process is controlled and capable, none of the items need to be inspected.*

The process capability is estimated to have limits.

$$\bar{y}_{\cdot} \pm 3s \qquad \text{or} \qquad \bar{y}_{\cdot\cdot} \pm 3\sqrt{MSE},$$

or the same formula incorporating bias adjustments for s or $\sqrt{MSE}$. For individuals, these are just the estimated control limits. An estimated process capability is also easily obtained from a means chart. The control limits in a means chart are $\mu \pm 3\sigma/\sqrt{N}$ where N is the group sample size. The process capability limits are $\mu \pm 3\sigma$. The difference between the two control limits is $6\sigma/\sqrt{N}$. Multiplying this difference of $6\sigma/\sqrt{N}$ by $\sqrt{N}/2$ gives 3σ from which it is easy to obtained the process capability. To illustrate the computations, reconsider the means chart in Figure 4.3. Recall that the blood pressure process is not under control, so the process has no capability. We are merely illustrating the computations for process capability. The difference between the control limits in Figure 4.3 is $95.15 - 83.72 = 11.43$. In Figure 4.3, $N = 4$, so the estimate of 3σ is $11.43(\sqrt{4}/2) = 11.43$. The estimated process capability would be the value of the center line, 89.43, plus and minus the estimate of 3σ. From Figure 4.3, the process capability would be estimated as 89.43 ± 11.43, i.e, the interval from 78.00 to 100.86 (if this process were under control and had a capability).

Figure 4.6 contains Minitab's process capability chart. As far as I can tell, this figure does not actually compare the capability limits to the specification limits. In the left box LSL and USL are the lower and upper specification limits, respectively, set here as blood pressures of 55 and 94. A target blood pressure was not specified. Below these in the left box are numbers related to the summaries from Table 4.2: the grand mean $\bar{y}_{\cdot\cdot}$, the total sample size $19 \times 4 = 76$, a bias adjusted $\sqrt{MSE}$ [StdDev(Within)], and the bias adjusted overall sample standard deviation computed from all 76 observations. (The bias adjustments converge to 1 when the sample sizes – technically the degrees

[2] To be predictable, a process need only be stationary, not iid. Positively correlated stationary processes would have artificially narrower control limits, hence a greater probability of getting outside the limits. BUT the justification for these control limits is not a probability statement but the experience that they work!

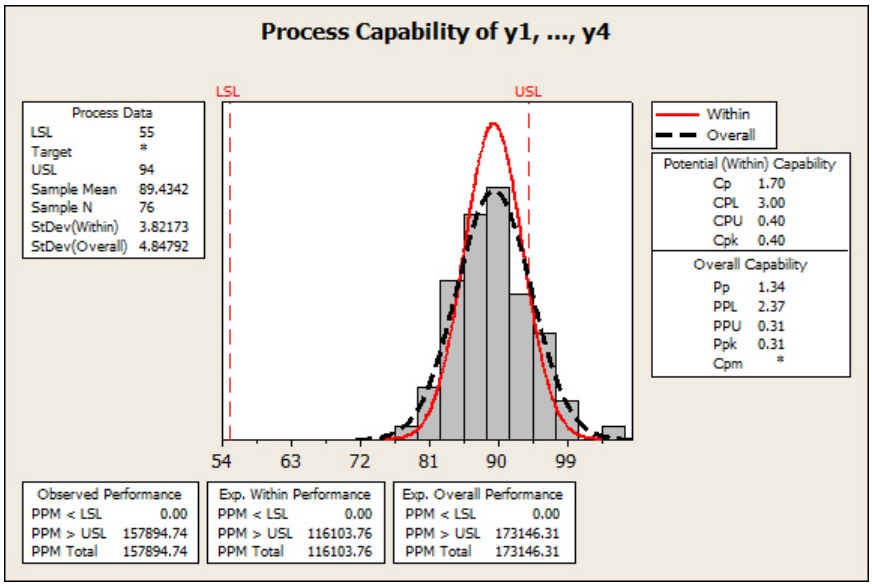

Figure 4.6: *Process Capability Chart for the Data of Table 4.2*

of freedom – get large and in this example the "overall" adjustment is negligible.) In the boxes below the graph, PPMs are parts per million (rather than percentages or estimated probabilities). The observed performance is just that. The expected within and overall performances are based on approximating normal distributions using the grand mean and an estimated standard deviation (either within or overall). The stuff in the right box are commonly used measures that I have not yet decided are worth caring about. For example, this C_p statistic is the length of the specification interval divided by the length of the process capability interval, so a value greater than 1 tells you that you could potentially meet specifications *if you were exactly on target*. The other "C" measures are refinements on C_p. (This C_p is **not** the commonly used Mallows' C_p created for variable selection in regression.)

4.2.1.1 Six-Sigma

There has been much discussion of achieving 6σ process control. There is a narrow focus of 6σ which is merely an aspect of process capability. There is also a general focus with Six-Sigma being the name of a quality management program, see Hahn et al. (1999) and the January 2002 issue of *Quality Progress*.

As an aspect of process capability, the 6σ concept has sometimes been misunderstood. When misunderstood, 6σ can easily lead to lowering the quality of products. For a process under control, the capability remains the $\mu \pm 3\sigma$ limits. Properly understood, 6σ *limits have been achieved when the specification interval contains the limits* $\mu \pm 6\sigma$. This can only happen when σ is very small relative to the variability allowed with the specification limits.

If the specification limits happen to be 3σ from the mean μ, so they happen to agree with the control limits, with normal data about 0.3% (3 out of 1000) units will be outside of specifications. If the variance has been reduced so that now the specification limits agree with $\mu \pm 6\sigma$, almost no units will be outside specifications. Anytime the $\mu \pm 6\sigma$ limits are within the specification limits your "process capability" is fantastic. (But you could potentially do even better in overall quality by moving μ closer to the specified target, say μ_0.)

The point is that 6σ is something to be achieved by reducing variability – which is very difficult. If 6σ is somehow imposed rather than being achieved, it has almost certainly been misunderstood

Table 4.3: *High Blood Pressure Readings*

Group (i)	N	# of High Pressures (y_i)	$\hat{p}_i$
1	25	13	0.52
2	25	10	0.40
3	25	9	0.36
4	25	3	0.12
5	25	2	0.08
6	25	2	0.08
7	25	2	0.08
8	25	2	0.08
9	25	1	0.04
10	25	0	0.00
11	25	1	0.04
12	25	1	0.04
13	25	3	0.12
14	25	2	0.08
15	25	11	0.44
16	25	4	0.16
17	25	1	0.04
18	25	1	0.04
Total	450	68	0.151

and the misunderstanding is ruining quality. For example, *making the incorrect claim that a process is under control whenever observations are within* $\mu \pm 6\sigma$ *will have disastrous results.*

As a method of quality management, Six-Sigma focuses on using quality methods to improve financial measures – something that is probably inevitable for a successful program given the nature of American business management.

4.3 Attribute Charts

Frequently, control charts are used to evaluate whether certain *attributes* of a process are under control. We now consider such charts.

EXAMPLE 4.3.1 *Np chart*

Diastolic blood pressures are considered to be high if they are 94 or larger. To illustrate what are known as *Np* charts, we dichotomized the 460 blood pressure observations into pressures that were high and those that were not. Further we divided the 460 observations into 18 consecutive groups of 25 while ignoring the last 10 observations. The data are given in Table 4.4. The values for $\hat{p}_i$ in the table are simply the observed proportions of high blood pressures, i.e., the number of high pressures divided by 25. An *Np* control chart is a plot of the number of high blood pressures versus the group. The chart is given in Figure 4.7.

If the 460 observations have the same distribution, there is some fixed probability p that an observation will be 94 or greater. If the observations are all independent, then the number of high pressures out of each group of 25 has a binomial distribution, i.e., the number of high counts y_i has $y_i \sim \text{Bin}(N, p)$ where $N = 25$ is the number of trials in each group. As seen in Section 3.4, $\text{E}(y_i) = Np$, $\text{Var}(y_i) = Np(1-p)$, and the standard deviation of y_i is $\sqrt{Np(1-p)}$. It follows that the center line for the *Np* chart is at Np and the control limits are at $Np \pm 3\sqrt{Np(1-p)}$. As mentioned, $N = 25$, but we do not know p and must estimate it. We combine information across groups. From Table 4.4 there are a total of 68 high blood pressure readings out of 450 total readings, so the total estimate of p is $68/450 = 0.151$. The center line is estimated as $25(0.151) = 3.778$. (Note that the center line value can also be computed as the number of high blood pressures averaged over the 18 groups, i.e., $68/18$.) The control limits are $3.778 \pm 3\sqrt{25(0.151)(1-0.151)}$. The number of high

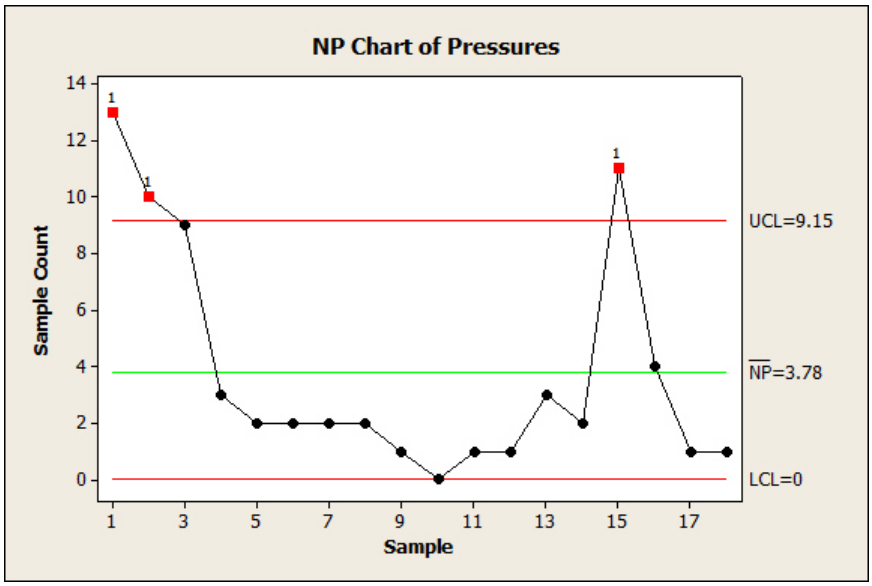

Figure 4.7: *Np Chart for High Diastolic Blood Pressures*

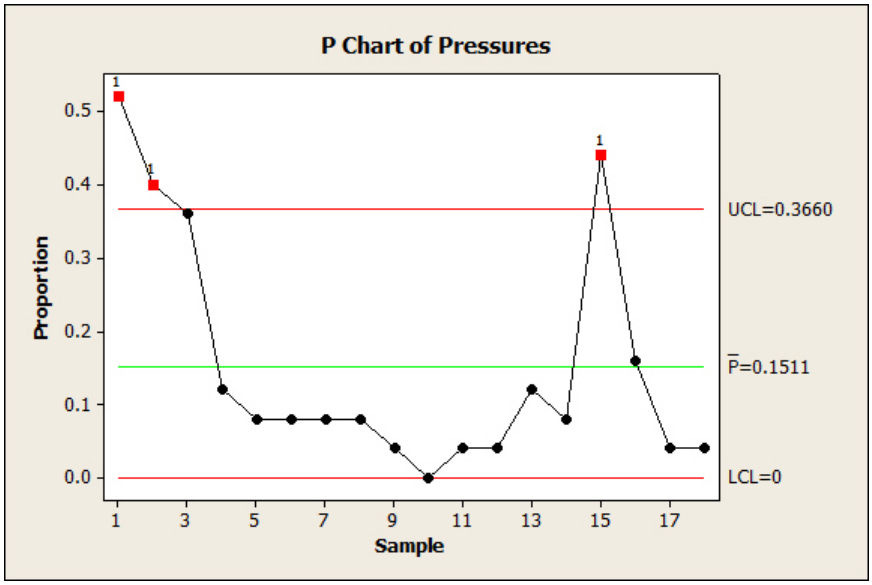

Figure 4.8: *p Chart for High Diastolic Blood Pressures*

blood pressures is nonnegative so if the computed lower control limit is negative, the limit is taken as 0. Similarly, the number of high blood pressures cannot be larger than $N = 25$.

From the *Np* chart, the process is obviously out of control. Three points are above the control limits while there is a string of 11 consecutive points that are below the center line. It should also be noted that, as with other charts, the attribute data for this chart was obtained by forming rational subgroups. We looked at the numbers of high readings in groups of 25 *consecutive* observations. And again, we have only 18 groups rather than Shewhart's suggested 25 so even if the process were not out of control, we would be hesitant to claim that it was in control. □

It is not always the case that one has the same number of trials in each group. The y_is could be $y_i \sim \text{Bin}(N_i, p)$ where the N_is change from group to group. In such cases the *Np* chart can be modified but the modified chart will not have a horizontal center line. In these cases it is convenient to form a *p* chart. In a *p* chart, the proportions (the $\hat{p}_i$s) are plotted against the group identifier. The center line of the plot is the pooled estimate of p, i.e., the total number of events divided by the total number of trials. For the high blood pressures, the $\hat{p}_i$s from Table 4.4 would be plotted against group number with a center line at $68/450 = 0.151$. In this plot, the upper and lower control limits may vary from group to group. For group i, $\hat{p}_i$ has $\text{E}(\hat{p}_i) = p$ and $\text{Var}(\hat{p}_i) = p(1-p)/N_i$. Thus the estimated control limits for the number of high blood pressures in a group of, say, $N_i = 10$ measurements are $0.151 \pm 3\sqrt{0.151(1-0.151)/10}$. Different groups with different numbers of trials N_i will have different control limits. For $N_i = 10$, the estimated upper limit is 0.49. The estimated lower limit is -0.19 which is adjusted to 0. If the proportion of high readings among 10 measurements is greater than 0.49, in other words if 5 or more of the 10 readings are high, the process is out of control.

EXAMPLE 4.3.2 *p chart*
Figure 4.8 gives the *p* chart for the data of Table 4.3. Because the N_is all equal 25, it looks just like the *Np* chart except all the entries are divided by 25. The following exercise adds more data and changes the results. □

EXERCISE 4.1 Shumway gave 460 BP readings. The first 450 are summarized in Table 4.3. The final 10 readings contain 0 high blood pressures. Reconstruct the *p* chart incorporating this extra group of data.

Example 4.3.4 also contains a *p* chart with substantial variation in the N_is.

The final types of control charts that we will discuss in any detail are *c* and *u* charts. These charts are based on count data having a Poisson distribution. Poisson data are similar to Binomial data with small probabilities of detecting an attribute but many trials in which to detect them. For example, the number of defects on the surface of a table might have a Poisson distribution where the probability of a defect at any location is small but there is no obvious limit to the number of locations on the table surface. If y_i is Poisson, and $\text{E}(y_i) = \mu$, then μ determines the entire distribution and in particular $\text{Var}(y_i) = \mu$.

The *c* chart involves plotting the number of occurrences versus the "group." (A group may be a single object like a table.) The center line is μ, which is identical to that in the *Np* chart. μ is estimated as the average number of defects per group. The control limits are $\mu \pm 3\sqrt{\mu}$ and are estimated by replacing μ with its estimate.

EXAMPLE 4.3.3 *c chart.*
We illustrate the *c* chart by once again using the data of Table 4.3. These data are more properly analyzed with an *Np* or *p* chart but the sample sizes are not small and the probability of getting a high blood pressure reading is not large, so the application is not too inappropriate. As in the *Np* chart, numbers of high readings are plotted against groups. As in the *Np* chart, the center line is 3.778. It is the average number of high readings per group, $68/18 = 3.778$. In the *c* chart, the control limits are $3.778 \pm 3\sqrt{3.778}$ as opposed to the *Np* chart values of $3.778 \pm 3\sqrt{3.778(1-0.151)}$. Even though the sample sizes of $N = 25$ are not terribly large and the estimated probability 0.151 is not terribly small, the *c* chart given in Figure 4.9 is very similar to the *Np* chart in Figure 4.7. The upper control limit is noticeably higher in Figure 4.9 but we still see three points above the UCL and 11 consecutive points below the center line. □

Variations on the *c* chart are also necessary. For example, when looking at the number of painting defects on a piece of sheet metal, if we have different sized pieces of sheet metal we need to adjust for that fact. These adjustments are made in what are known as *u charts*. Suppose the observations y_i

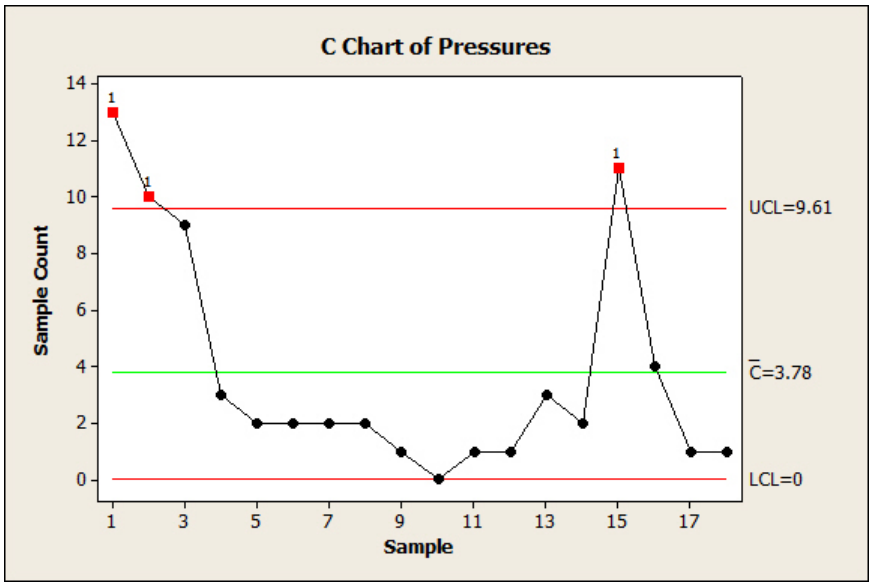

Figure 4.9: *c Chart for High Diastolic Blood Pressures*

Table 4.4: *Moisture Resistance Test Failures*

Lot	Size	Failures	Lot	Size	Failures
1	80	3	14	200	27
2	125	7	15	200	1
3	80	4	16	200	9
4	80	7	17	125	0
5	200	0	18	200	1
6	80	0	19	200	1
7	200	9	20	315	5
8	125	1	21	315	2
9	125	0	22	200	0
10	125	0	23	50	1
11	125	8	24	200	2
12	125	1	25	315	3
13	200	1	26	315	3

are Poisson with mean $\lambda\ell_i$ where λ is a failure rate and ℓ_i is a measure of size for the ith observation. The chart is based on plotting the values y_i/ℓ_i. The mean of these values is λ and the variance is λ/ℓ_i, so the chart is based on $\lambda \pm 3\sqrt{\lambda/\ell_i}$. Here λ is estimated from the data, not as the mean of the y_i/ℓ_i values, but as

$$\hat{\lambda} \equiv \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \ell_i}.$$

EXAMPLE 4.3.4 *u chart.*
King (1995, qe95-495) presents data on the number of failures per lot observed during moisture resistance tests on a plastic-encapsulated metal-film resistor. These are given in Table 4.4. Note that the sample sizes vary from lot to lot, so a p chart is appropriate, cf. Figure 4.10. Also notice that all of the sample sizes are quite large and the rates of defects are small, so the p chart will be well approximated by a u chart, cf. Figure 4.11. Also see Exercises 4.9.18 and 4.9.22.

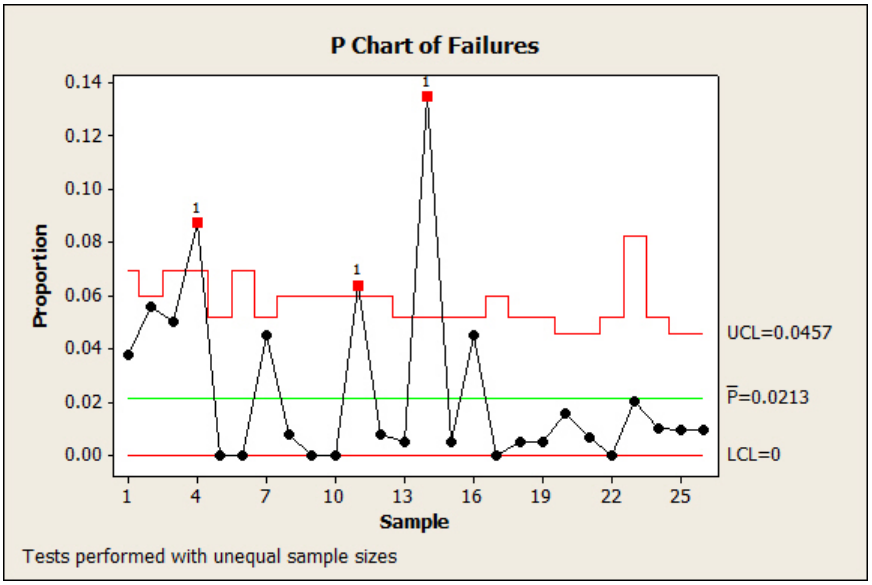

Figure 4.10: *p Chart for moisture resistance failures.*

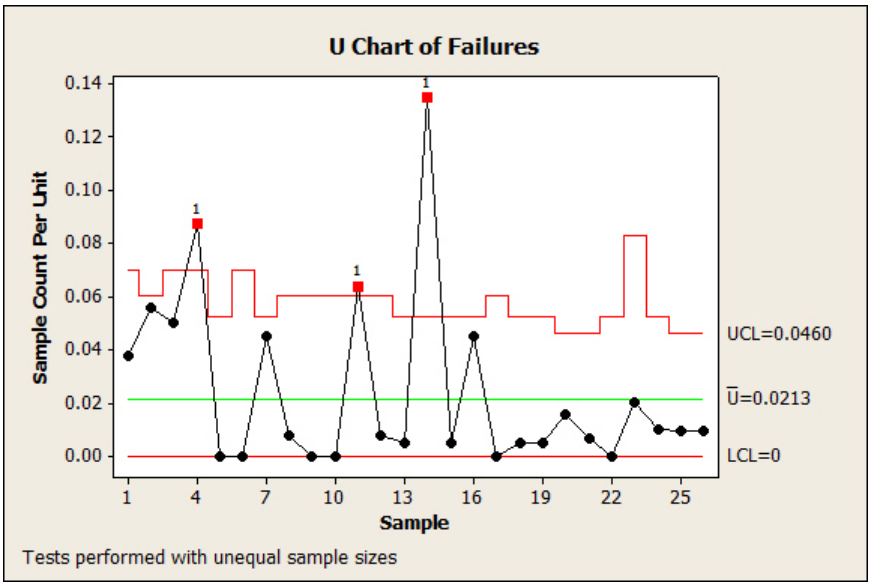

Figure 4.11: *u Chart for moisture resistance failures.*

4.4 Control Chart Summary

Suppose we have data on a process y_h, $h = 1, \ldots, n$. *If the process is under control*, any manner of selecting the data, as long as the selection does not depend on the y_h values, should provide a random sample from the process. In other words, the y_hs should be independent and identically distributed (*iid*). Unfortunately, we cannot look at a group of numbers and tell whether they are observations from an iid process. The point of control charts is to create an operational definition of what it means to be iid. If the control chart is satisfied, then the data are close enough to iid for us to use them for reliable prediction.

Control charts constitute an **operational definition**, not a statistical test, but like a statistical test, they are set up assuming things about the data, in this case that the data are iid. The idea is that if the data are not iid, the control charts have a good chance of spotting the problem. We assume that for $h = 1, \ldots, n$,

$$\mathrm{E}(y_h) = \mu, \qquad \mathrm{Var}(y_h) = \sigma^2.$$

Most often μ and σ^2 are two unrelated, unknown parameters but in some problems (for attribute data) σ^2 is actually a function of μ.

If the process is under control (i.e., if the observations are iid) then it should remain (largely) within the interval having the (control) limits

$$\mu \pm 3\sigma.$$

These limits also define the *process capability*. If the process capability interval is contained within the specification limits associated with the product, we are good to go. *A process that is out of control has no capability!!* The "Six-Sigma" program is named for the requirement that $\mu \pm 6\sigma$ be within the specification limits. Such a requirement necessitates both reduced variability and (usually) that the target μ be near the center of the specification limits.

A key idea, perhaps *the* key idea, behind control charts is dividing the data into rational subgroups having essentially identical conditions. When that is done, it becomes convenient to replace the subscript h with a pair of subscripts ij wherein i identifies the rational subgroup and j identifies observations with the group. Rewrite the data as y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. Here we must have a total sample size of $N_1 + \cdots + N_n$. As before,

$$\mathrm{E}(y_{ij}) = \mu, \qquad \mathrm{Var}(y_{ij}) = \sigma^2.$$

As discussed earlier, Shewhart recommends $n \geq 25$.

From within each subgroup, we define a statistic of interest, say $\hat{\theta}_i \equiv \hat{\theta}_i(y_{i1}, \ldots, y_{iN_i})$. These are typically the sample mean, sample standard deviation, or perhaps the sample range. The statistic of interest is chosen so that the expected value is the same for every group, say,

$$\mathrm{E}(\hat{\theta}_i) = \theta$$

but the variance may depend on the group size, so write

$$\mathrm{Var}(\hat{\theta}_i) = V_i.$$

A theoretical control chart would plot the pairs $(i, \hat{\theta}_i)$, with a central horizontal line at θ and control limits

$$\theta \pm 3\sqrt{V_i}.$$

In many charts, V_i is the same for all i, which makes for a more attractive chart. In practice, we need to estimate both θ and the V_is.

As discussed in Section 2, in addition to requiring that the plotted points remain within the control limits, there are a number of other criteria that one may incorporate into the operational definition of a process being under control.

Table 4.5: *Control Chart Summary.* σ *estimated by* $\sqrt{MSE}$*, i.e., not bias corrected.*

Chart	$\hat{\theta}_i$	θ	$\hat{\theta}$	$\sqrt{V_i}$	$\sqrt{\hat{V}_i}$	Notes
Means	$\bar{y}_{i\cdot}$	μ	$\bar{y}_{\cdot\cdot}$	$\sqrt{\sigma^2/N_i}$	$\sqrt{MSE/N_i}$	
Variance	s_i^2	σ^2	MSE	$\frac{2\sigma^2}{N_i-1}$	$\frac{2MSE}{N_i-1}$	
S	s_i	$c_{N-1}\sigma$	$c_{N-1}\sqrt{MSE}$	$\sqrt{C_{N-1}\sigma^2}$	$\sqrt{C_{N-1}MSE}$	$N \equiv N_1 = \cdots = N_n$
R	R_i	$d_N\sigma$	$\bar{R}$	$\sqrt{D_N\sigma^2}$	$\sqrt{D_N MSE}$	$N \equiv N_1 = \cdots = N_n$
NP	$y_i \equiv N\bar{y}_{i\cdot}$	$Np \equiv n\mu$	$N\hat{p} = N\bar{y}_{\cdot\cdot}$	$\sqrt{Np(1-p)}$	$\sqrt{N\hat{p}(1-\hat{p})}$	$N \equiv N_1 = \cdots = N_n$
P	$\hat{p}_i \equiv \bar{y}_{i\cdot}$	$p \equiv \mu$	$\hat{p} = \bar{y}_{\cdot\cdot}$	$\sqrt{\frac{p(1-p)}{N_i}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{N_i}}$	
C	y_i	μ	$\hat{\mu} \equiv \sum y_i/n \equiv \bar{y}_{\cdot}$	$]sqrt\mu$	$\sqrt{\hat{\mu}}$	$y_i \sim \text{Pois}(\mu)$
U	y_i/ℓ_i	λ	$\hat{\lambda} \equiv \sum y_i/\sum \ell_i$	$\sqrt{\lambda/\ell_i}$	$\sqrt{\hat{\lambda}/\ell_i}$	$y_i \sim \text{Pois}(\lambda\ell_i)$

Table 4.5 summarizes the standard control charts within this context. In the Np and p attribute charts we assume that the y_{ij}s only take on the values 0 and 1. The U and C charts essentially have $N_i = \infty$ but that causes no problem. The dispersion/variability charts all incorporate values that are only valid for data with normal distributions. In particular, the use of the χ^2 distribution and computing the bias constants c_{N_i-1}, C_{N_i-1}, d_{N_i}, and D_{N_i} all involve assuming the y_{ij} have normal distributions. With $(N_i-1)s_i^2/\sigma^2 \sim \chi^2(N_i-1)$, write c_{N_i-1} as the constant for which $\text{E}(s_i) = c_{N_i-1}\sigma$. Write C_{N_i-1} for the value with $\text{Var}(s_i) = C_{N_i-1}\sigma^2$. Write d_{N_i} for the value such that $\text{E}(R_i) = d_{N_i}\sigma$ and D_{N_i} for the value with $\text{Var}(R_i) = D_{N_i}\sigma^2$. I am happy to estimate σ with $\sqrt{MSE}$ and have done so in the table. Some people, including the makers of Minitab, prefer to estimate σ with $\sqrt{MSE}/c_{\sum N_i - n}$, so one could change the table accordingly. Also, when $N \equiv N_1 = \cdots = N_n$ one could alternatively use $\bar{R}/d_N$ as the estimate of σ.

As discussed in Chapter 6, a process that generates iid data is not the only process for which reliable predictions can be made. If the process is *stationary*, the associated data will also be predictable. Informally, a stationary processes is one in which, no matter when you choose to start looking at the process, future and past observations always show the same patterns. For this to be true, the mean and variance cannot change with time, but the observations do not have to be independent. In practice we can only evaluate past data, but we can look at "future" data relative to some point in the past. An iid process is always stationary, but stationary processes need not be iid. It is much more difficult to give an operational definition for a general stationary process than for an iid process. One needs to be very careful if trying to claim that a process that fails control charts is still predictable because the control chart failure is due to the process being stationary but not iid.

4.5 Average Run Lengths

When testing statistical hypotheses, tests are often evaluated by their size and power. The size is the probability of rejecting an hypothesis when it is true and the power is the probability of rejecting an hypothesis when it is false. The average run length is a related concept for control charts. The run length is the (random) number of points you plot until you see a point that is out of control. The average run length is the expected value of the random run length. We want the average run length to be large for processes that are under control and short for processes that are out of control. The EWMA and CUSUM charts in Section 7 were developed to have shorter average run lengths for out of control processes.

4.6 Discussion

Control charts are really designed as an operational definition of what it means for a process to be under control. Not infrequently, statisticians want to interpret them as statistical tests for whether the mean of the process has jumped from its historical/previous value or for whether the data display trends, i.e., whether the mean of the process is gradually increasing or decreasing. Two useful procedures for conducting such tests are CUSUM (cumulative sum) charts and EWMA (exponentially weighted moving average) charts and are discussed in the next section.

In a statistical test one makes a series of assumptions and checks to see whether the data are consistent with those assumptions. Most often, a particular assumption is isolated as the null hypothesis and if the data are inconsistent with the assumptions we conclude that the null hypothesis is false. Logically, concluding that the null hypothesis is false requires us to believe that the other assumptions are valid.

The whole point of control charting is to decide whether the data are independent observations from a fixed process (population). It is not difficult to jump to the conclusion that this is an assumption we wish to test. However, one can take a very different philosophical view. One can take the view that the problem is not one of testing the assumptions but rather one of defining the terms used in the assumptions. What does it actually mean for, say, 50 observations to be independent from the same population. If we have an *operational definition* of what these terms mean, we will be able to look at 50 numbers and decide whether they are independent observations from the same population or not. The philosophy behind control charts is not to test ill defined assumptions but rather to give an operational definition of statistical control, see Shewhart (1939, p. 40). A person looks at various characteristics of a control chart and if they are appropriate one concludes that the process satisfies the operational definition of being in statistical control. This is not too different from the philosophy frequently adopted in Statistics. Typically, we do an informal check of mathematical assumptions and, if the assumptions are not too bad, we proceed with formal statistical inferences, i.e, tests, confidence intervals, and prediction intervals. The only real difference is that in control charting, the checks on mathematical assumptions are formalized into an operational definition while the inferences are informal prediction intervals. It is of interest to note that typically informal checks on assumptions do not include methods for checking independence, whereas we will see in Subsection 9.1.1 that means charts (and, it turns out, Np, p, c, and u charts) are effective at detecting lack of independence when rational subgroups have been chosen.

As pointed out above, many of the characteristics of traditional control charts (bias adjustments, variances of dispersion measures) are calculated by using results that relate to independent normally distributed data. It follows that normal distributions are being built into the operational definitions of statistical control whereas normality is not part of the mathematical definition of statistical control. In our earlier discussion we were somewhat critical of this feature. From the philosophical point of view presented here, our earlier discussions about the validity of assumptions need to be recast as discussions on the nature of an appropriate operational definition for statistical control.

Deming (1986, Chapter 11) alludes to these issues. On pages 334 and 335 he seems to suggest that one should just use control charts without worrying about whether the data satisfy the assumptions underlying the calculations. But earlier, on page 319, he indicates that usable decision rules must be made in advance and that the 3σ control limits work well "under a wide range of unknowable circumstances". In fact, I believe he is saying that one needs operational definitions and that the 3σ control limits have been found to provide a useful definition. More generally, Deming argues that a major problem in economic activities is that management fails to provide workers with operational definitions and consequently, workers never know whether they have done a task successfully. Shewhart (1939) discusses these issues in more detail.

4.7 Testing Mean Shifts from a Target

Control charts (with or without the additional criteria discussed by Nelson, 1984) are used to create an operational definition for what it means to have a process under control. We now present two sequential statistical tests for whether a process is on target at a specified value μ_0. Only if you let the data empirically determine the value μ_0 can these be considered as potential additional criteria for whether a process is under control, because being under control has nothing to do with being on target.

Exponentially weighted moving average (EWMA) charts and Cumulative Sum (CUSUM) charts provide sequential tests for whether a process experiences a mean shift away from a target. We will discuss the tests in terms of a sequence of observations $y_1, y_2, y_3, \ldots, y_n$ that are independent and have the same variance σ^2. Let $\mathrm{E}(y_i) = \mu_i$. The tests assume a null hypothesis that the means remain stable throughout the process at a fixed target value $\mathrm{E}(y_i) = \mu_0$. These procedures are good at picking up the alternative that, up to some time T, the mean remains on target $\mathrm{E}(y_i) = \mu_0$, $i = 1, \ldots, T$ but then the entire process shifts slightly off target so that now $\mathrm{E}(y_i) = \mu_0 + \delta$, $i = T+1, \ldots, n$ for some small $\delta \neq 0$. They work best when T is small. The process control versions of these tests are more sensitive than individuals (or means) charts when δ is relatively small.

Both procedures can be used for process control using the empirical choice $\mu_0 = \bar{y}_\cdot$, similar in spirit to the eight tests discussed by Nelson (1984). In fact, except at the beginning of the data sequence, EWMA charts do not depend heavily on the value μ_0, so if you get past the first few observations an EWMA chart can be used as a criterion for process control even without empirically choosing μ_0.

When applied to individual observations y_i, the variance σ^2 is estimated as it would be in an individuals chart. However both procedures can also be applied to a sequence of mean values $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \bar{y}_{3\cdot}, \ldots$ obtained from rational subgroups of size N. In this case the variance of an "observation" σ^2 is replaced by $\mathrm{Var}(\bar{y}_{i\cdot}) = \sigma^2/N$ and σ^2 is well estimated by pooling the within group sample variances.

4.7.1 Exponentially Weighted Moving Average Charts

Exponentially weighted moving average (EWMA) charts are a tool often used alongside control charts but they are really tests for whether the data display a mean shift away from a target μ_0. The chart uses the specified center line μ_0 and plots the EWMA over time while checking whether the EWMA remains within 3 standard deviations of the center line.

For the purpose of constructing an EWMA chart, an EWMA is commonly defined via the recursive relationship

$$\tilde{\mu}_t \equiv (1-\alpha) y_t + \alpha \tilde{\mu}_{t-1},$$

where $0 < \alpha < 1$ and $\tilde{\mu}_0 \equiv \mu_0$. It is by no means clear what this has to do with an exponentially weighted moving average.

The obvious way to define an EWMA is to apply exponentially decreasing weights to each observation, i.e., for $0 < \alpha < 1$,

$$\hat{\mu}_t \equiv (\alpha y_t + \alpha^2 y_{t-1} + \alpha^3 y_{t-2} + \cdots + \alpha^t y_1) \Big/ \sum_{i=1}^{t} \alpha^i = \frac{\sum_{i=1}^{t} \alpha^i y_{t-i+1}}{\sum_{i=1}^{t} \alpha^i}.$$

The point of this is that older observations get multiplied by much smaller numbers. We will later show that when t is large, $\hat{\mu}_t$ very nearly satisfies the recursive definition for $\tilde{\mu}_t$ and we will give explicit formulae for $\tilde{\mu}_t$ that display exponential downweighting of older observations.

As will be shown later, the standard deviation for $\tilde{\mu}_t$ with a fixed starting point μ_0 is

$$\sigma \sqrt{\left(\frac{1-\alpha}{1+\alpha}\right)(1-\alpha^{2t})}.$$

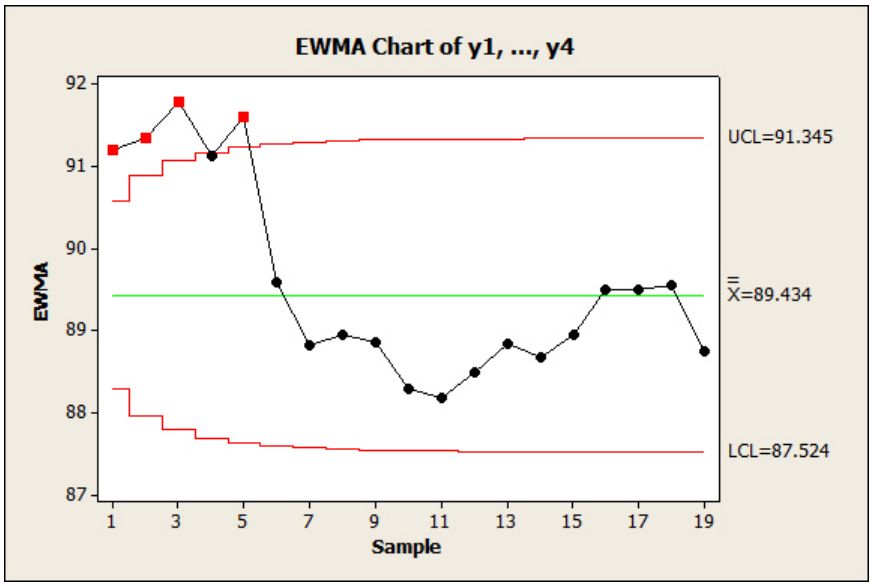

Figure 4.12: *Exponentially weighted moving average chart for grouped diastolic blood pressures.*

Typically, one just rejects the constant mean hypothesis whenever a plotted value gets outside of 3 estimated standard deviations from the target μ_0. The plot displays all of these results for $t = 1,2,3,\ldots,n$. Note that as t gets large, the standard deviation rapidly approaches the limit

$$\sigma\sqrt{\left(\frac{1-\alpha}{1+\alpha}\right)}.$$

The variance σ^2 is estimated as in the individual charts of Section 4.1 or, when appropriate, the means charts of Section 4.2. Lucas and Saccucci (1990) facilitate the construction of more formal tests.

Of course one needs to specify the value of α. A common default choice is $\alpha = 0.8$. Many discussions of EWMA charts are presented in terms of

$$\lambda \equiv 1-\alpha,$$

in which case the common default choice is $\lambda = 0.2$.

To turn the "on target" test of $H_0 : \mu_i = \mu_0$, $i = 1,\ldots,n$ into an under control test of $H_0 : \mu_i = \mu$, $i = 1,\ldots,n$ for come unspecified μ, we use $\mu_0 = \bar{y}_\cdot$, the empirical choice. When doing so, people typically use the same variance formula ignoring the fact that $\bar{y}_\cdot$ is now random. One obvious alternative choice is to start the recursion with $\tilde{\mu}_1 \equiv y_1$, for which the variance will be given in the next subsection.

To a great extent, the niceties of turning the "on target" EWMA test into an under control criterion can be ignored because the exponential weighting soon discounts the target value μ_0 out of the picture. For the first few steps t, the procedure is sensitive to whether $\mathrm{E}(y_t) = \mu_0$, but before long the chart essentially reduces to a test of whether there is a constant mean for all observations or whether at some point in time the mean shifts.

Figure 4.12 gives the empirically centered EWMA chart with $\alpha = 0.8$ for the grouped blood pressure data of Table 4.2. Because it is grouped data the value of σ in the formulae should be replaced with $\sqrt{MSE/N}$ or possibly a bias corrected version of that.

4.7.1.1 *Derivations of results*

Three fundamental equalities about power series are needed. The first is obtained by looking at each term in the product $(1-\alpha)\sum_{i=0}^{\infty}\alpha^i$,

$$\frac{1}{1-\alpha} = \sum_{i=0}^{\infty}\alpha^i, \qquad \frac{\alpha}{1-\alpha} = \sum_{i=1}^{\infty}\alpha^i,$$

and finally,

$$\begin{aligned}
\sum_{i=1}^{t}\alpha^i &= \sum_{i=1}^{\infty}\alpha^i - \sum_{i=t+1}^{\infty}\alpha^i \\
&= \sum_{i=1}^{\infty}\alpha^i - \alpha^{t+1}\sum_{i=0}^{\infty}\alpha^i \\
&= \frac{\alpha}{1-\alpha} - \alpha^{t+1}\frac{1}{1-\alpha} \\
&= \frac{\alpha-\alpha^{t+1}}{1-\alpha} \\
&\doteq \frac{\alpha}{1-\alpha},
\end{aligned}$$

when t is large.

We can now see that $\hat{\mu}_t$ approximately satisfies the recursive relationship that defines $\tilde{\mu}_t$ when t is large.

$$\begin{aligned}
\hat{\mu}_t &\equiv \frac{\sum_{i=1}^{t}\alpha^i y_{t-i+1}}{\sum_{i=1}^{t}\alpha^i} \\
&= \frac{\alpha y_t + \sum_{i=2}^{t}\alpha^i y_{t-i+1}}{(\alpha-\alpha^{t+1})/(1-\alpha)} \\
&\doteq \frac{\alpha y_t + \sum_{i=2}^{t}\alpha^i y_{t-i+1}}{\alpha/(1-\alpha)} \\
&= (1-\alpha)y_t + \frac{\sum_{i=2}^{t}\alpha^i y_{t-i+1}}{\alpha/(1-\alpha)} \\
&= (1-\alpha)y_t + \alpha\frac{\sum_{i=1}^{t}\alpha^i y_{t-1-i+1}}{\alpha/(1-\alpha)} \\
&\doteq (1-\alpha)y_t + \alpha\frac{\sum_{i=1}^{t-1}\alpha^i y_{(t-1)-i+1}}{\alpha-\alpha^t/(1-\alpha)} \\
&= (1-\alpha)y_t + \alpha\hat{\mu}_{t-1}.
\end{aligned}$$

Assuming the data are uncorrelated (or independent) with $\mathrm{E}(y_t)=\mu$ and $\mathrm{Var}(y_t)=\sigma^2$,

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_t) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{t}\alpha^i y_{t-i+1}}{(\alpha-\alpha^{t+1})/(1-\alpha)}\right) \\
&= \frac{\sum_{i=1}^{t}\alpha^{2i}\mathrm{Var}(y_{t-i+1})}{[(\alpha-\alpha^{t+1})/(1-\alpha)]^2} \\
&= \sigma^2\frac{\sum_{i=1}^{t}\alpha^{2i}}{[(\alpha-\alpha^{t+1})/(1-\alpha)]^2} \\
&= \sigma^2\frac{(\alpha^2-\alpha^{2t+2})/(1-\alpha^2)}{[(\alpha-\alpha^{t+1})/(1-\alpha)]^2}
\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 \left(\frac{\alpha^2 - \alpha^{2t+2}}{(\alpha - \alpha^{t+1})^2} \right) \left(\frac{(1-\alpha)^2}{1-\alpha^2} \right) \\
&= \sigma^2 \frac{\alpha^2 - \alpha^{2t+2}}{(\alpha - \alpha^{t+1})^2} \left(\frac{1-\alpha}{1+\alpha} \right) \\
&= \sigma^2 \left(\frac{1-\alpha^{2t}}{(1-\alpha^t)^2} \right) \left(\frac{1-\alpha}{1+\alpha} \right) \\
&= \sigma^2 \left(\frac{1+\alpha^t}{1-\alpha^t} \right) \left(\frac{1-\alpha}{1+\alpha} \right) \\
&\doteq \sigma^2 (1-\alpha)/(1+\alpha).
\end{aligned}
$$

When t is small it pays to use the exact formula.

From the recursive definition of $\tilde{\mu}_t$ that starts with $\tilde{\mu}_1 = y_1$,

$$
\begin{aligned}
\tilde{\mu}_t &= (1-\alpha) y_t + \alpha \tilde{\mu}_{t-1} \\
&= (1-\alpha) y_t + \alpha \left[(1-\alpha) y_{t-1} + \alpha \tilde{\mu}_{t-2} \right] \\
&= (1-\alpha) y_t + \alpha (1-\alpha) y_{t-1} + \alpha^2 \tilde{\mu}_{t-2} \\
&= (1-\alpha) y_t + \alpha (1-\alpha) y_{t-1} + \alpha^2 \left[(1-\alpha) y_{t-2} + \alpha \tilde{\mu}_{t-3} \right] \\
&= (1-\alpha) y_t + \alpha (1-\alpha) y_{t-1} + \alpha^2 (1-\alpha) y_{t-2} + \alpha^3 \tilde{\mu}_{t-3} \\
&= (1-\alpha) y_t + \alpha (1-\alpha) y_{t-1} + \alpha^2 (1-\alpha) y_{t-2} + \alpha^3 \left[(1-\alpha) y_{t-3} + \alpha \tilde{\mu}_{t-4} \right] \\
&= (1-\alpha) \sum_{i=0}^{t-2} \alpha^i y_{t-i} + \alpha^{t-1} \tilde{\mu}_1 \\
&= (1-\alpha) \sum_{i=0}^{t-2} \alpha^i y_{t-i} + \alpha^{t-1} y_1.
\end{aligned}
$$

This involves exponential weighting but not as clearly as $\hat{\mu}$. Again assuming the data are uncorrelated with $\mathrm{E}(y_t) = \mu$ and $\mathrm{Var}(y_t) = \sigma^2$,

$$
\begin{aligned}
\mathrm{Var}(\tilde{\mu}_t) &= \mathrm{Var} \left((1-\alpha) \sum_{i=0}^{t-2} \alpha^i y_{t-i} + \alpha^{t-1} y_1 \right) \\
&= (1-\alpha)^2 \sum_{i=0}^{t-2} \alpha^{2i} \mathrm{Var}(y_{t-i}) + \alpha^{2(t-1)} \mathrm{Var}(y_1) \\
&= \sigma^2 \left[(1-\alpha)^2 \sum_{i=0}^{t-2} \alpha^{2i} + \alpha^{2(t-1)} \right] \\
&= \sigma^2 \left[\frac{(1-\alpha)^2}{\alpha^2} \sum_{i=1}^{t-1} \alpha^{2i} + \alpha^{2(t-1)} \right] \\
&= \sigma^2 \left[\frac{(1-\alpha)^2}{\alpha^2} \frac{\alpha^2 - \alpha^{2t}}{1-\alpha^2} + \alpha^{2(t-1)} \right] \\
&= \sigma^2 \left[\left(\frac{(1-\alpha)^2}{1-\alpha^2} \right) \left(\frac{\alpha^2 - \alpha^{2t}}{\alpha^2} \right) + \alpha^{2(t-1)} \right] \\
&= \sigma^2 \left[\left(\frac{1-\alpha}{1+\alpha} \right) (1 - \alpha^{2(t-1)}) + \alpha^{2(t-1)} \right].
\end{aligned}
$$

For the more common choice of a nonrandom starting point for the recursive definition of $\tilde{\mu}_t$, say $\tilde{\mu}_0 = \mu_0$, we get

$$
\tilde{\mu}_t = (1-\alpha) \sum_{i=0}^{t-1} \alpha^i y_{t-i} + \alpha^t \mu_0.
$$

and

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}_t) &= \mathrm{Var}\left((1-\alpha)\sum_{i=0}^{t-1}\alpha^i y_{t-i} + \alpha^t \tilde{\mu}_0\right)\\ &= (1-\alpha)^2\sum_{i=0}^{t-1}\alpha^{2i}\mathrm{Var}(y_{t-i})\\ &= \sigma^2\left(\frac{1-\alpha}{1+\alpha}\right)(1-\alpha^{2t}).\end{aligned}$$

4.7.2 CUSUM charts

CUSUM charts again involve a specified target μ_0 that is being tested. The basic CUSUM statistic is

$$C_t = \sum_{i=1}^{t}(y_i - \mu_0).$$

The standard deviation of C_t is $\sqrt{t}\sigma$ so if the process is on target it should remain within $\pm 3\sqrt{t}\sigma$. A plot would display the three numbers C_t, $\pm 3\sqrt{t}\sigma$ for $t = 1,2,3,\ldots,n$. *But why make life that simple?* The additional complications of standard procedures are related to their theory as sequential statistical tests. (For just about any conceivable sample size the $\pm 3\sqrt{t}\sigma$ limits are wider than the *Law of the Iterated Logarithm* limits $\pm\sigma\sqrt{2t\log(\log(t))}$ that for arbitrarily large t contain all of the observations with probability one, so clearly there is room for improvement.)

Standard CUSUM procedures have upper and lower control limits of $\pm h\sigma$ and require another *slack parameter* k such that mean shifts up to $\pm k\sigma$ are not considered worth bothering about. Thus, a standard CUSUM will be good for detecting small shifts off of μ_0 but only small shifts that are at least $\pm k\sigma$. Quite a bit of work can go into selecting the values h and k, cf. Woodall and Adams (1993). Minitab uses the defaults $h = 4$ and $k = 0.5$. The `qcc` R package uses the defaults $h = 5$ and $k = 1$.

We now define upper and lower (high and low) cumulative sums that respectively look for a drift of the process above the target and a drift of the process below the target. A process is identified as off target if either of the following numbers gets outside the control limits,

$$C_t^h = \max\{0, C_{t-1}^h + (y_t - \mu_0) - k\sigma\}$$

and

$$C_t^l = \min\{0, C_{t-1}^l + (y_t - \mu_0) + k\sigma\}.$$

Here $C_0^h = C_0^l = 0$. (Starting at points other than 0 is sometimes used to give a "fast initial response.") As long as the cumulative sum C_t remains larger than $kt\sigma$, C_t^h is positive, but, once the upper cumulative sum goes negative, C_t^h zeros out and starts again looking for a positive drift. As long as $C_i \geq ki\sigma$, $i = 1,\ldots,t-1$, the process is declared off target when $C_t > (h+kt)\sigma$. *This initially seems unreasonably stringent (typical default values for $(h+tk)\sigma$ are much larger than $3\sqrt{t}\sigma$) but the high and low CUSUMs are quite easy to zero out which restarts the the process of checking for a mean shift and restarting seems like a good thing.* (It is similar to the downweighting of earlier observations used in the EWMA.) This procedure is sometimes called the tabular procedure as opposed to George Barnard's more old-fashioned V-mask procedure. [The use of the *High and Low* cumulative sums was well illustrated in the quality training film of the same name by Akira Kurasawa.]

Replacing μ_0 with $\bar{y}_{\cdot}$ provides a method for identifying an out of control process. The variance is estimated as in an individuals chart when plotting individual observations y_i or as in a means chart when plotting rational subgroup means $\bar{y}_{i\cdot}$. Figure 4.13 applies the default Minitab CUSUM procedure to the grouped blood pressure data.

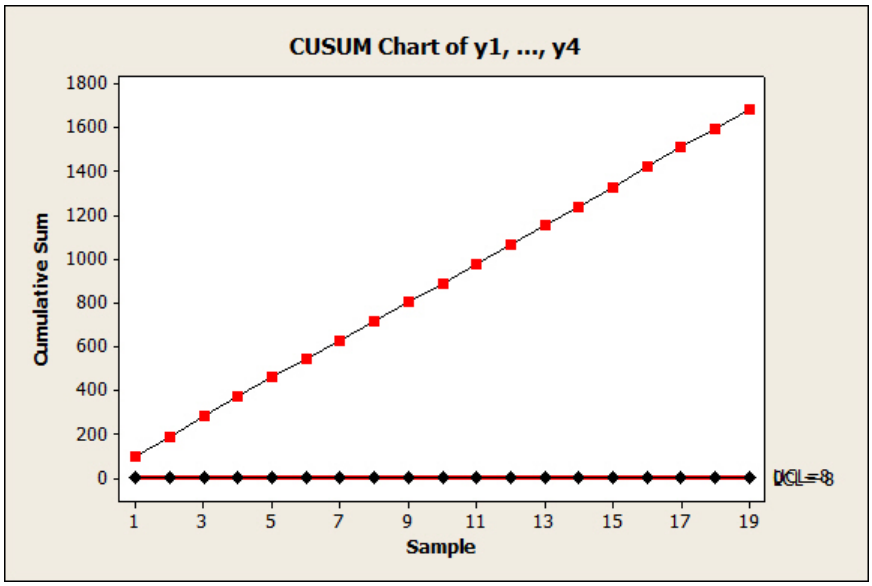

Figure 4.13 *Cusum chart for diastolic blood pressures. The default is* $\mu_0 = 0$ *which, given the vertical scale, is why* $\pm h\sigma/\sqrt{N}$ *and* C_t^l *are indistinguishable on this plot.*

4.8 Computing

4.8.1 Minitab

The whole point of using Minitab is that it is really easy to use. It is menu driven, the menus are easy to use, the menus construct the underlying code, and even the underlying code is very easy to use. Minitab went through 19 versions before they decided to run it off the web so that they can continually improve it. For academic access, see `estore.onthehub.com`.

The Minitab menus are along the top. The current version displays five windows. On the left is a Navigator window for selecting the output to look at in the the Session window at the center top. The center bottom displays the Worksheet containing the data. The worksheet operates rather like a spreadsheet. On the top right is a Command Line window for entering and running commands. The bottom right is a History window of the commands run. If you use the menu structure, the History window shows you the commands that the menus generated.

Minitab is not good at reading my data files. Before reading any file you should look at it and compare it to the corresponding table in the book. (Notepad and Wordpad are good Windows editors for doing this.) To read my data files, count the number of columns of data and remove any variable names. The file for Table 4.2 is `tab4-2.dat` and it contains variable names, so I removed them and saved the data as `tab4-2a.dat`. The file has 9 columns so use the commands

```
read c1-c9;
file "c:\e-drive\books\quality\bookdata\tab4-2a.dat".
```

My "path" is

```
e-drive\books\quality\bookdata
```

Yours will be different. *If you are copying commands from a* `.pdf` *file, some characters do not always copy appropriately, like ~ and -. You may need to delete and retype them.*

Small data files can also be copied and pasted directly into the worksheet. After copying the data, put your cursor in the worksheet at the top left of where you want to paste the the data. The menu that appears when you try to `Paste` is pretty self explanatory.

Generally when reading files one would go to the `File` menu on the top left and choose `Open`. In

the new window find the appropriate folder and file and open it. Check the preview and, for my data files, change the Field delimiter option to Space. Typically, uncheck the Data has column names box but you can see whether to do that from the preview. If the preview looks ok, hit OK. *On my data files this rarely works* because I included extra spaces to make them look good in an editor.

The following Minitab commands were used to obtain Figure 4.1 from data file tab4-1.dat. *The data file looks a little funky but it is just one column of numbers* and it can be read using the standard Minitab menus. For the current version of Minitab, copy the following text into the Command Line window on the top right, highlight it, and hit the "Run" button at the bottom of that window.

```
describe c1
ichart c1;
mu 92.880;
sigma 3.799.
```

Minitab will show you the ichart output. To see the describe output go to the "navigator" window on the left and click on "descriptive statistics." Semicolons are used to let Minitab know you are specifying a subcommand. The final period indicates that you are done with a command that involves subcommands. In fact, I had to run the describe command prior to the ichart command because I used the describe output to get the values I put into the mu and sigma subcommands of ichart.

Older versions of Minitab provided prompts of MTB > or SUBC> prior to commands

```
MTB > describe c1
MTB > ichart c1;
SUBC> mu 92.880;
SUBC> sigma 3.799.
```

To reproduce the Individuals chart Figure 4.1 using the menus, after reading the data of Table 4.1, go to the Stat menu and then the Describe option to compute the mean and standard deviation. Then, again within the stat menu, choose the Control Charts option and choose Variables Charts for Individuals, choose Individuals and enter the mean and standard deviation using the I Chart Options button. Of course you could just run the individuals chart without entering the mean and standard deviation but it will look a bit different.

The following Minitab commands were used to generate the control charts for Example 4.2.1 from file tab4-2.dat. The xbar command allows subcommands for specifying μ and σ and for specifying various control chart tests. This application uses the subcommand RSub to specify that the rational subgroup data are in columns c2 through c5, each row of these columns being a different group. The commands for performing *s* and *R* charts work similarly and the menus do a good job of explaining the options.

```
xbar;
RSub c2-c5;
mu 89.434;
sigma 3.81;
tests 1:8.
schart;
RSub c2-c5.
rchart;
RSub c2-c5;
rbar.
```

An alternative data structure would be to have a single column of numbers, say c15, with the rational subgroups being each, say, 4 consecutive numbers in c15.

```
xbar c15 4;
mu 89.434;
sigma 3.81;
tests 1:8.
```

4.8.2 R Commands

R has lots of relevant capabilities including the package/library qcc for basic charts. Start by looking at https://cran.r-project.org/web/packages/qcc/vignettes/qcc_a_quick_tour.html. If you want to use R you might want to look at http://www.stat.unm.edu/~fletcher/Rcode.pdf which gives a short introduction to R and covers most, but not all, of the regression, analysis of variance, and experimental design topics considered here. Also, http://www.stat.unm.edu/~fletcher/R-TD.pdf gives some, and over time should give more, code for topics covered in TiD. At the moment it mostly directs you to relevant R packages for experimental design.

I have not actually run any of this code yet. The code reads in nine columns (vectors) including y1,y2,y3,y4 for which row i gives the four observations on group i. The object s.data (denoting Shumway data), as used below, requires y1,y2,y3,y4 to be rearranged into one vector of numbers y with a corresponding index vector i to identify its groups. There may be more efficient ways to define s.data from y1,y2,y3,y4.

```
rm(list = ls())
bp <- read.table("C:\\E-drive\\Books\\quality\\BookData\\tab4-2a.dat",sep="",
         col.names=c("Group","y1","y2","y3","y4","N","ybar-i","s-i","Range"))
attach(bp)
bp

# Rearrange y1,y2,y3,y4 into y as indicated
# HAVEN'T DONE IT YET

#install.packages("qcc")
library(qcc)

sdata=qcc.groups(y,i)
qcc(sdata,type="xbar")
qcc(sdata,type="S")
qcc(sdata,type="R")

# If the group structure in y is every 4 rows
sdata=y
qcc(sdata,type="xbar",size=4)
qcc(sdata,type="S",size=4)
qcc(sdata,type="R",size=4)
```

The package qcc also has capabilities for process capability, the various attribute charts, EWMA and CUSUM charts, as well as the multivariate charts of Appendix A and the Pareto charts and cause-and-effect diagrams of Chapter 2.

```
mycc=qcc(sdata,type="xbar")
process.capability(mycc,c(spec1,spect2))

qcc(y,sizes=N,type="Np")
qcc(y,sizes=N,type="p")
qcc(y,type="c")
```

Table 4.6: *Ball Bearing Weights*

16.5	10.9	10.6	11.2	11.4	10.2	13.4
10.2	12.7	12.6	12.8	12.0	16.5	9.7
10.1	11.3	12.5	13.9	9.7	8.9	9.9
10.1	8.8	12.0	8.7	14.2	12.3	10.7
10.4	13.2	9.7	12.2	11.0	10.6	10.3
12.3	11.4	10.7	13.4	10.3	10.5	11.8
10.2	15.4	8.7	10.6	12.6	10.0	10.8
15.4	13.1	11.4	7.6	12.5	13.0	13.0
13.5	13.1	7.5	11.1			

Read across and then down.

```
qcc(y,sizes=N,type="u")

ewma()
cusum()

mqcc

pareto.chart()
cause.and.effect()
```

4.9 Exercises

EXERCISE 4.9.1 Construct an individuals chart from the data on ball bearing weights in Table 4.6. Which, if any, bearings are out of control?

EXERCISE 4.9.2 Using consecutive groups of size four, construct s and R charts from the ball bearing data of Table 4.6. Which, if any, points are out of control on the s chart? Which, if any, points are out of control on the R chart? Do you detect any difference in the sensitivity of the two charts?

EXERCISE 4.9.3 Construct a means chart from the ball bearing data of Table 4.6 using groups of four. Which, if any, points are out of control?

EXERCISE 4.9.4 What is the capability of the process explored in Table 4.6. The target value for the ball bearings is 12.5. How well is the manufacturer doing?

EXERCISE 4.9.5 Construct an individuals chart from the data of Table 4.7 on diameters of a coupling made for a refrigerator. Which, if any, couplings are out of control?

EXERCISE 4.9.6 Using consecutive groups of size four, construct s and R charts from the data of Table 4.7. Which, if any, points are out of control on the s chart? Which, if any, points are out of control on the R chart? Do you detect any difference in the sensitivity of the two charts?

EXERCISE 4.9.7 Construct a means chart from the data of Table 4.7 using groups of four. Which, if any, points are out of control?

EXERCISE 4.9.8 What is the capability of the process explored in Table 4.7. The target value for the coupling diameters is 22. How well is the manufacturer doing?

Table 4.7: *Coupling Diameters*

21.8	23.5	21.9	22.2	21.8	22.9	21.7	21.7
24.7	21.1	21.7	22.6	20.7	22.6	23.4	22.6
20.8	22.1	22.5	21.8	22.5	21.3	22.1	21.1
22.6	21.8	21.6	21.6	21.1	19.4	27.3	20.4
22.6	22.0	22.3	21.1	22.3	21.9	21.9	18.2
22.9	21.3	22.6	20.7	22.6	20.4	22.7	22.2
22.8	21.5	22.6	20.6	22.6	21.9	22.4	22.7
19.1	23.2	22.8	23.3				
Read across and then down.							

Table 4.8: *Side Board Lengths*

Group	Mean	Variance	Group	Mean	Variance	Group	Mean	Variance
1	19.7	0.97	10	16.6	2.83	19	16.5	9.41
2	18.5	4.77	11	15.8	6.71	20	16.1	1.25
3	14.7	0.32	12	14.3	7.59	21	19.7	2.85
4	13.3	5.56	13	15.2	4.64	22	20.7	3.51
5	15.0	5.84	14	18.7	2.49	23	16.6	0.91
6	16.4	1.71	15	15.8	4.92	24	17.2	9.93
7	18.1	0.28	16	14.4	6.04	25	20.3	3.25
8	13.9	2.43	17	17.4	0.54			
9	17.8	0.37	18	17.2	13.09			

EXERCISE 4.9.9 Table 4.8 contains sample means and variances for rational subgroups of four units observing the lengths of a side board used in the construction of a television stand. Construct a means chart and an *s* chart and evaluate whether the process is under control.

EXERCISE 4.9.10 Table 4.9 contains sample means and standard deviations (S. D.) for rational subgroups of four units. The measurements pertain to the length of the top of a television stand. Construct a means chart and an *s* chart and evaluate whether the process is under control.

EXERCISE 4.9.11 Table 4.10 contains the number of defectives in groups of 40 ball bearings. Construct an *Np* chart from the data and evaluate the state of control of the process.

EXERCISE 4.9.12 Table 4.11 contains the number of defectives in groups of 30 refrigerator couplings. Construct an *Np* chart from the data and evaluate the state of control of the process.

EXERCISE 4.9.13 Table 4.12 contains the number of defectives in groups of various sizes of ball bearings. Construct a *p* chart from the data and evaluate the state of control of the process.

Table 4.9: *TV Stand Top Lengths*

Group	Mean	S. D.	Group	Mean	S. D.	Group	Mean	S. D.
1	33.9	2.77	10	37.1	1.37	19	33.9	2.52
2	39.8	0.66	11	30.8	3.24	20	36.7	2.66
3	37.5	1.23	12	35.8	3.76	21	32.0	6.04
4	32.4	0.67	13	35.5	1.63	22	35.7	5.31
5	38.6	2.07	14	34.6	1.82	23	29.8	7.29
6	35.8	5.21	15	31.1	1.53	24	33.2	5.19
7	39.5	2.20	16	41.3	3.18	25	33.3	9.54
8	35.5	2.05	17	30.8	2.59			
9	33.1	1.38	18	33.3	1.65			

Table 4.10: *Defective Ball Bearings*

2	3	3	4	2	1	2	3	5	2
2	0	2	5	3	5	2	1	3	4
2	2	5	3	4	2	6	3	5	2
Read across and then down.									

Table 4.11: *Defective Refrigerator Couplings*

2	5	4	2	2	3	4	3	7	4
4	4	4	3	5	3	1	5	4	3
10	4	8	3	3	9	5	3	7	13
1	2	4	4	5	6	7	5	6	8
Read across and then down.									

EXERCISE 4.9.14 Construct a c chart for the data of Table 4.10 and compare the results to the Np chart computed in Exercise 4.9.11.

EXERCISE 4.9.15 Data on the purity of a product were provided by van Nuland (1994) and are reproduced in Table 4.13. Construct a means chart and a dispersion chart to evaluate whether the process is under control.

EXERCISE 4.9.16 *Hopper Data.*
The data in Table 4.14 were provided by Schneider and Pruett (1994). They were interested in whether the measurement system for the weight of railroad hopper cars was under control. A standard hopper car weighing about 266,000 pounds was used to obtain the first 3 weighings of the day on 20 days. The process was to move car onto the scales, weigh the car, move car off, move car on, weigh the care, move it off, move it on, and weigh it a third time. The tabled values are the weight of the car minus 260,000. Summary statistics are given in Table 4.15.

EXERCISE 4.9.17 *Injection Molding Data.*
Schneider and Pruett (1994) present data on the outside diameters of bottles made by an injection molding machine. The machine has four heads, so four bottles can be made at once. We treat these as rational subgroups. The data are given in Table 4.16 with summary statistics in Table 4.17. Create a means chart and a dispersion chart and draw conclusions from these charts. There is a problem with these data that the means and dispersion charts do not pick up. We will explore the problem further in Chapter 9. Check to see if there are too many observations too close to the center line of the control chart.

Table 4.12: *Defective Ball Bearings*

Group	Size	Defectives	Group	Size	Defectives	Group	Size	Defectives
1	40	3	11	40	2	21	20	4
2	40	1	12	40	3	22	20	2
3	40	2	13	40	1	23	20	3
4	40	3	14	40	7	24	20	2
5	40	2	15	40	3	25	20	3
6	40	5	16	40	3	26	20	1
7	40	2	17	40	3	27	20	2
8	40	3	18	40	5	28	20	3
9	40	4	19	40	6	29	20	3
10	40	3	20	40	4	30	20	5

Table 4.13: *Purity of a product*

1	8.76	8.81	9.21	9.00
2	9.13	9.18	9.01	9.35
3	9.34	9.16	8.91	8.98
4	9.14	9.12	9.13	9.14
5	9.21	8.58	8.53	9.14
6	8.55	8.79	9.04	8.88
7	9.09	8.85	8.90	8.67
8	8.81	8.88	8.68	9.17
9	9.31	8.85	9.51	8.93
10	9.14	9.15	9.06	8.85
11	8.92	8.57	8.62	9.50
12	9.33	8.64	9.23	9.20
13	8.97	9.33	9.22	8.86
14	9.03	9.07	8.85	8.72
15	9.29	8.97	8.75	9.33
16	9.28	8.96	9.26	8.88
17	9.05	9.13	8.93	8.85
18	9.19	9.06	9.45	9.07
19	8.51	9.10	9.52	9.25
20	8.56	8.64	8.85	9.51
21	9.17	9.27	8.89	9.15
22	9.04	8.63	9.49	8.85
23	8.69	9.00	9.10	9.25
24	8.72	9.18	9.06	9.23
25	8.71	9.52	9.85	9.81

Table 4.14: *Multiple weighings of a hopper car*

Day	First	Second	Third	Day	First	Second	Third
1	5952	5944	6004	11	5986	5920	5944
2	5930	5873	5895	12	6036	6084	6054
3	6105	6113	6101	13	6035	6136	6128
4	5943	5878	5931	14	6070	6016	6111
5	6031	6009	6000	15	6015	5990	5950
6	6064	6030	6070	16	6049	5988	6000
7	6093	6129	6154	17	6139	6153	6151
8	5963	5978	5966	18	6077	6012	6005
9	5982	6005	5970	19	5932	5899	5944
10	6052	6046	6029	20	6115	6087	6078

Table 4.15: *Summary Statistics for Hopper Data*

DAY	N	MEAN	STDEV	DAY	N	MEAN	STDEV
1	3	5966.7	32.6	11	3	5950.0	33.4
2	3	5899.3	28.7	12	3	6058.0	24.2
3	3	6106.3	6.1	13	3	6099.7	56.1
4	3	5917.3	34.6	14	3	6065.7	47.6
5	3	6013.3	15.9	15	3	5985.0	32.8
6	3	6054.7	21.6	16	3	6012.3	32.3
7	3	6125.3	30.7	17	3	6147.7	7.6
8	3	5969.0	7.9	18	3	6031.3	39.7
9	3	5985.7	17.8	19	3	5925.0	23.3
10	3	6042.3	11.9	20	3	6093.3	19.3

Table 4.16: *Outside diameter of injection molded bottles*

	Head					Head			
Sample	1	2	3	4	Sample	1	2	3	4
1	2.01	2.08	2.08	2.04	11	1.99	1.99	2.09	2.11
2	1.97	2.03	2.09	2.10	12	1.98	2.02	2.03	2.08
3	2.03	2.09	2.08	2.07	13	1.99	1.98	2.05	2.04
4	1.96	2.06	2.07	2.11	14	2.01	2.05	2.07	2.08
5	1.94	2.02	2.06	2.11	15	2.00	2.05	2.06	2.06
6	2.01	2.03	2.07	2.11	16	2.00	2.00	2.08	2.14
7	2.00	2.04	2.09	2.06	17	2.01	2.00	2.05	2.15
8	2.01	2.08	2.09	2.09	18	2.03	2.09	2.11	2.12
9	2.01	2.00	2.02	2.07	19	1.99	2.10	2.09	2.09
10	2.01	1.96	2.08	2.11	20	2.01	2.01	2.01	2.11

Table 4.17: *Summary Statistics for Injection Data*

Sample	N	MEAN	STDEV	Sample	N	MEAN	STDEV
1	4	2.0525	0.0340	11	4	2.0450	0.0640
2	4	2.0475	0.0602	12	4	2.0275	0.0411
3	4	2.0675	0.0263	13	4	2.0150	0.0351
4	4	2.0500	0.0638	14	4	2.0525	0.0310
5	4	2.0325	0.0718	15	4	2.0425	0.0287
6	4	2.0550	0.0443	16	4	2.0550	0.0681
7	4	2.0475	0.0377	17	4	2.0525	0.0685
8	4	2.0675	0.0386	18	4	2.0875	0.0403
9	4	2.0250	0.0311	19	4	2.0675	0.0519
10	4	2.0400	0.0678	20	4	2.0350	0.0500

EXERCISE 4.9.18 *p and u charts.*
Construct the *p* and *u* charts for Example 4.3.4.

EXERCISE 4.9.19 Clark and Milligan (1994, qe94-389) report the data in Table 4.18 on filling barrels of honey that were provided by Dutch Gold Honey for use in making Haagen-Daz special blend. The goal was to fill the barrels with at least 660 units, but as little more than that as possible. The values reported are the decimal values *x* for readings that are 660.*x* units. Construct a means chart and a dispersion chart and evaluate the results.

EXERCISE 4.9.20 Table 4.19 gives data from van Nuland (1992, qe96-444) and Chao and Cheng (1996) on a DSC apparatus. Construct a means chart and a dispersion chart for these data and evaluate the results.

Table 4.18: *Haagen-Daz special blend*

	Time				Time		
Sample	1	2	3	Sample	1	2	3
1	0	4	2	11	1	2	2
2	2	4	1	12	2	2	1
3	2	4	1	13	1	3	2
4	0	2	2	14	4	2	2
5	3	0	1	15	4	0	2
6	4	4	0	16	0	2	1
7	2	2	3	17	2	2	4
8	1	4	2	18	0	4	2
9	0	0	4	19	4	3	1
10	0	0	2				

Table 4.19: *Data on a DSC apparatus*

Group	Observations			
1	2.2	2.3	2.0	2.1
2	1.9	2.1	1.8	1.6
3	1.7	1.8	1.8	1.9
4	1.5	2.0	1.7	2.0
5	1.7	2.2	2.2	1.5
6	2.2	1.5	2.1	1.8
7	1.8	2.1	1.8	1.9
8	2.4	1.8	1.9	2.1
9	2.2	1.9	2.5	2.8
10	1.8	2.6	1.4	2.3
11	2.4	2.0	1.8	2.1
12	1.0	1.1	1.4	1.5
13	2.0	1.4	0.7	0.4
14	0.9	1.1	1.3	0.7
15	1.1	0.9	1.0	1.3
16	1.0	1.6	1.3	0.9
17	1.7	1.7	1.0	1.0
18	1.3	1.0	0.9	1.7
19	1.4	1.0	0.9	1.6
10	0.9	0.5	0.7	1.5
21	0.9	1.5	1.8	1.7
22	1.8	0.9	1.5	0.6
23	1.2	1.9	2.1	1.7
24	1.5	1.7	1.5	1.0
25	1.5	1.0	1.4	1.2

Table 4.20: *Automobile Assembly Data*

−0.2302	0.9196	1.2809	0.9288	0.0430	0.4074	1.9086	0.1009
1.1807	−0.1985	0.2714	−0.1458	−0.1269	0.0350	1.5164	1.1236
−0.0814	−0.3408	2.1323	0.3683	0.1120	0.3759	−0.2727	0.2610
1.2097	0.3638	1.3688	2.0566	0.5649	0.7998	0.9407	0.3447
0.8107	0.1849	0.4069	0.2707	0.0701	−0.0498	−0.6768	0.5909
0.3270	0.3255	0.7791	1.5059	0.4731	1.1318	0.5256	0.8239
−0.1833	1.3530	−0.2356	−0.9623	0.8188	−0.0955	−0.3808	0.3889
0.7145	0.2056						
Read across and then down.							

EXERCISE 4.9.21 Table 4.20 gives Radson and Alwan (1996, qe96-171) data from a car assembly plant. Laser sensors are used to measure auto dimensions, and these data are a key dimension for attaching front wheel components. Proper placement of these components is related to vehicle stability when driving the car. The measurements are deviations relative to the standard given in millimeters. Give an individuals chart for these data and evaluate the results.

EXERCISE 4.9.21 When constructing a u chart we use the estimatate

$$\hat{\lambda} \equiv \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \ell_i}$$

rather than more obvious estimate

$$\tilde{\lambda} \equiv \frac{\sum_{i=1}^{n} y_i/\ell_i}{n}.$$

Show that

$$\text{Var}\left(\hat{\lambda}\right) \leq \text{Var}\left(\tilde{\lambda}\right).$$

Chapter 5

Prediction from Other Variables

The primary purpose of Statistics is to allow us to make predictions. Indeed, prediction is a primary aspect of all science and technology. If we turn on a light switch, we do so in anticipation that our little corner of the world will become brighter. When a school looks at the results of a student's college preparatory exam, they do so because they believe the results help predict the performance of the student in college. Turning on a light switch actually causes the world to get brighter, but a college preparatory exam does not cause a student to do well or poorly in school. The college preparatory exam is merely correlated with student performance. High exam scores tend to go with good performance and low exam scores with poor performance — but the relationship is far from perfect.

The point of control charts is that if a process is under control, we can expect the future behavior of the process to be like the past behavior that we observed. In this chapter we examine the use of one variable, such as preparatory exam score, to predict another variable, such as grade point average (gpa) after two semesters. To make such predictions, we will need training data on the exam scores and 2nd semester gpas for a collection of students. From these data we can develop a prediction model and when a new student is presented to us with their exam score, we can predict their 2nd semester gpa. An important aspect of data based prediction is that the data for the prediction must be collected in a similar fashion to the training data. For example, a predictive model based entirely on data from suburban high school students will probably not work well for predicting the performance of inner city high school students.

Not all prediction models are created equally. Although ACT and SAT scores unquestionably have some ability to predict 2nd semester gpa, there is a lot of error in such predictions. On the other hand, if you are rolling a smooth round steel ball over a given smooth surface for exactly 2 seconds, the initial velocity of the ball will be very good at predicting how far the ball will roll. But again, the quality of the predictions depend on having the future observations being taken similarly to the past observations. If we change the steel ball or the surface on which it is rolling, using the past data to predict these new events is highly questionable.

Statistical analysis cannot properly be used as the basis for inferring causation. Statistical analysis is only concerned with correlation, i.e., the tendency of two events to occur together. We have already discussed the fact that college preparatory exam scores do not in any reasonable way cause 2nd semester gpas. We said that turning on a light switch causes more light, but the causation that we infer is not based on statistics. It is based on the idea that turning the switch will complete an electrical circuit that will cause the light bulb to work. The statistics involved are simply the data that the switch is turned and the light comes on. It could be that turning on the switch gives an electrical jolt to a rat who then often decides to eat something and the rat's food tray is set up so that the light goes on when the rat eats something. In this case the rat eating can be considered a cause, but the electrical jolt given the rat is merely correlated with the light coming on — no matter how reliable is the event of the light coming on when the switch is thrown.

Table 5.1: *Prater's gasoline–crude oil data*

y	x	y	x	y	x	y	x
6.9	235	10.0	267	24.8	360	32.1	444
14.4	307	15.2	300	26.0	365	34.7	345
7.4	212	26.8	367	34.9	395	31.7	402
8.5	365	14.0	351	18.2	272	33.6	410
8.0	218	14.7	379	23.2	424	30.4	340
2.8	235	6.4	275	18.0	428	26.6	347
5.0	285	17.6	365	13.1	273	27.8	416
12.2	205	22.3	275	16.1	358	45.7	407

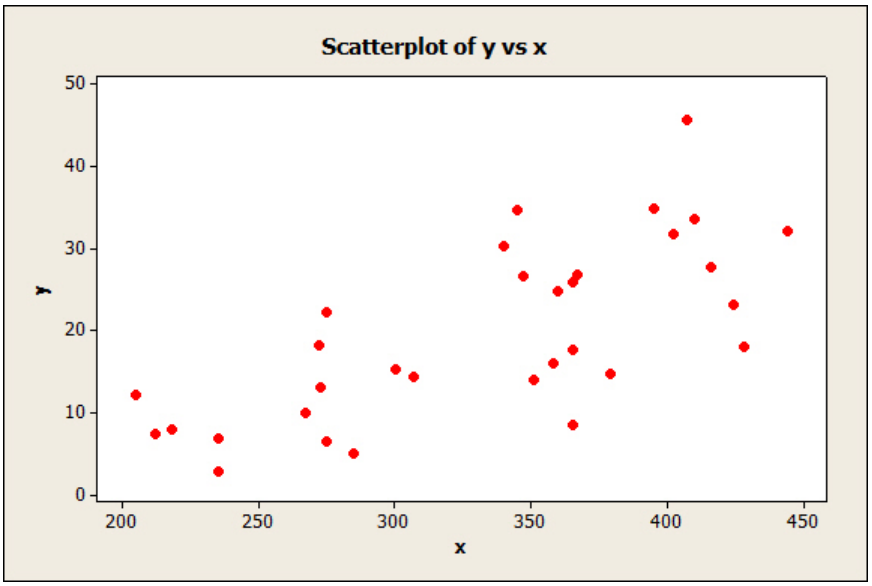

Figure 5.1: *Gasoline percentage versus temperature.*

5.1 Scatter Plots and Correlation

A scatter plot is a simple tool for seeing whether a relationship exists between two variables and it is even a simple tool for making predictions.

EXAMPLE 5.1.1. Table 5.1 gives data from Hader and Grandage (1958), Atkinson (1985), and Christensen (1996, Exercise 13.8.3.) on y, the percent of gasoline obtained from crude oil,and x, the temperature in °F at which all the crude oil is vaporized. Figure 5.1 is a scatter plot showing the relationship between x and y. From the plot, if the temperature were 325°F, we would predict about 15% gasoline. More realistically, we might predict somewhere between 9% and 24%. Notice that the data seem to be more variable as the temperature increases, e.g., the data are more spread out around 425°F than they are around 225°F.

In the next section we will introduce a mathematical model for prediction. The model is going to assume that the variability in the data does not depend on the temperature. Since that is not the case for these data, we need to do something. A standard method of dealing with this problem is to try transforming the data. In Figures 5.2, 5.3, and 5.4, we have plotted three standard transformations of the data: $\sqrt{y}$, $\log(y)$, and $1/y$ versus temperature. In Figure 5.2 the data seem to maintain the same variability throughout. In Figures 5.3 and 5.4, the data get less variable as temperature increases.

From Figure 5.2 we might predict the square root of the gasoline percentage for 325°F to be about 4 with an interval of about 2.5 to 5. This corresponds to predictions of 16% with an interval

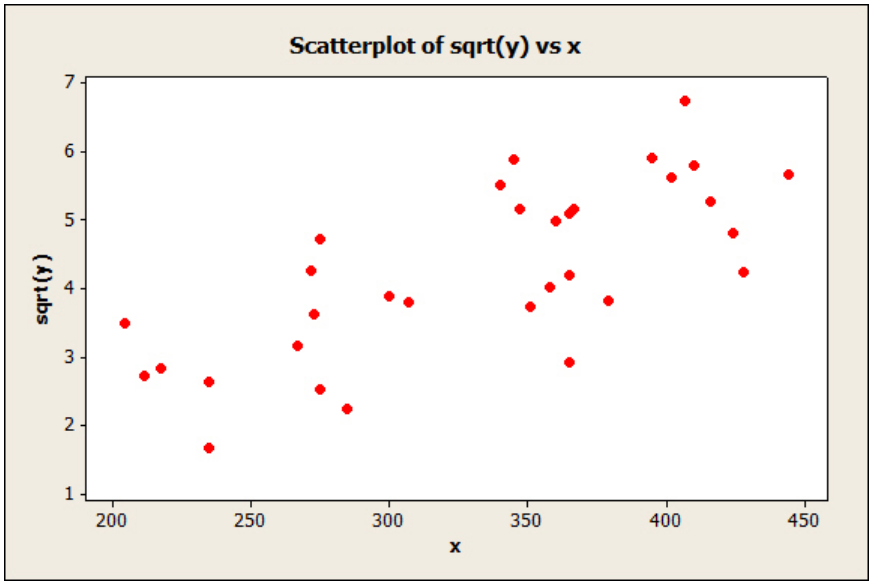

Figure 5.2: *Square root of gasoline percentage versus temperature.*

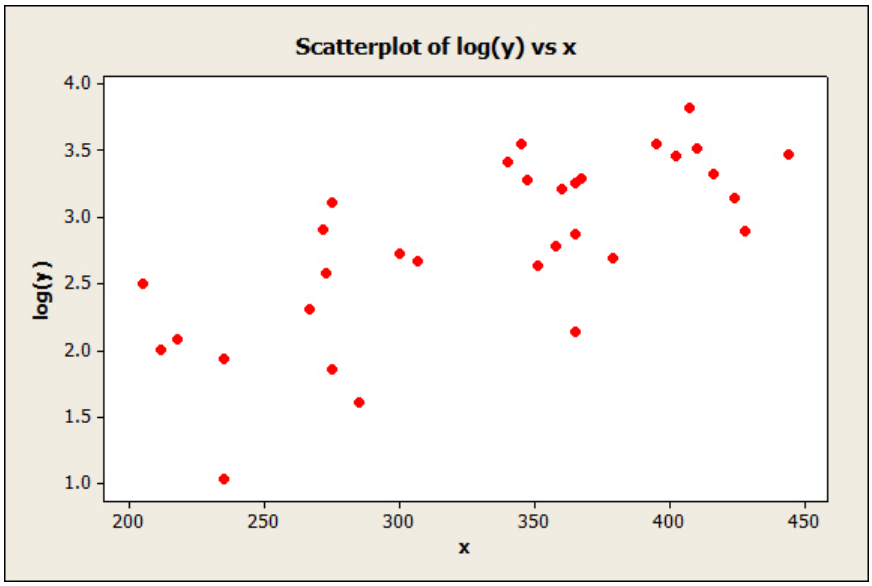

Figure 5.3: *Natural logarithm of gasoline percentage versus temperature.*

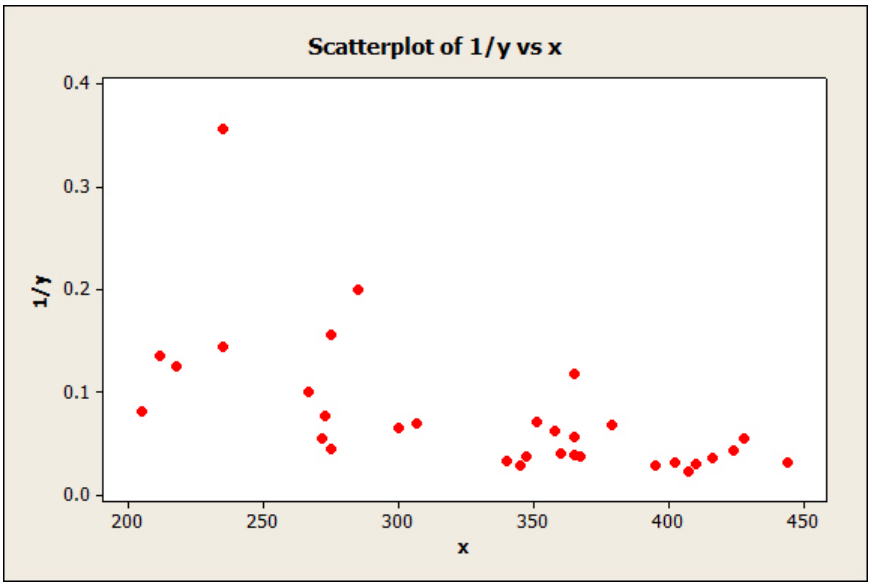

Figure 5.4: *Reciprocal of gasoline percentage versus temperature.*

of 9% to 25%. Admittedly, these are very crude and come from simply looking at the picture. They are also about what we got from Figure 5.1. For some purposes, these predictions may be entirely adequate. In the next section, we develop more sophisticated prediction methods.

The correlation coefficient is a unitless number between -1 and 1 that measures the linear relationship between two variables. A value of 0 indicates no linear relationship and values of -1 and 1 indicate perfect linear relationships. A value of 1 indicates a perfect positive linear relationship. For example if the correlation between temperature and square root of gasoline percentage was 1, then the plot of the two variables would be a perfect line with a positive slope. The positive slope indicates that as one variable increases, the other also increases. In Figure 5.2, the data form a line but it is far from perfect. In particular, the sample correlation is the sample covariance divided by the product of the sample standard deviations:

$$r \equiv \frac{s_{xy}}{s_x s_y} = 0.729.$$

Similarly, a value of -1 indicates a perfect negative linear relationship. If the correlation between two variables is -1, then the plot of the two variables would be a perfect line with a negative slope, so that as one variable gets larger, the other variable gets smaller.

5.2 Simple Linear Regression

What we are seeing in both Figures 5.1 and 5.2 is something that looks basically like a line. However, it is not a perfect linear relationship, so the line must incorporate some error. For n data points, we use the statistical model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{5.2.1}$$

$i = 1, \ldots, n$ where the line is $y = \beta_0 + \beta_1 x$ and the errors ε_i are assumed to be independent and identically distributed as $N(0, \sigma^2)$. In this model, y_i is an observable random variable and x_i is treated as a fixed known number. There are three unknown parameters, β_0 and β_1 from the line and σ^2, the variance of the errors. We estimate β_0 and β_1 by taking them to minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Notice that the y_is and x_is are observed, so only β_0 and β_1 are unknown. The minimizing values, say $\hat{\beta}_0$ and $\hat{\beta}_1$ are called least squares estimates. It turns out that

$$\hat{\beta}_0 = \bar{y}_\cdot - \hat{\beta}_1 \bar{x}_\cdot; \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}.$$

The variance σ^2 is estimated by something call the mean squared error (MSE). The estimated errors (residuals) are

$$\hat{\varepsilon}_i \equiv y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

and we average their squared values to get the MSE,

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Note that in averaging the squared errors, we divide by $n-2$ rather than n, that is because to obtain the estimated errors we had to estimate two parameters β_0 and β_1. MSE is an unbiased estimate of σ^2, i.e.,

$$\mathrm{E}(MSE) = \sigma^2.$$

Similarly, it turns out that

$$\mathrm{E}(\hat{\beta}_0) = \beta_0; \qquad \mathrm{E}(\hat{\beta}_1) = \beta_1.$$

The least squares estimates have

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{(n-1)s_x^2}\right]; \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(n-1)s_x^2}.$$

Therefore,

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}_\cdot^2}{(n-1)s_x^2}\right]}; \qquad \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{MSE}{(n-1)s_x^2}}.$$

5.2.1 Basic Analysis

For the gasoline data of Table 5.1, the estimated regression equation is

$$\sqrt{\hat{y}} = -0.137 + 0.0132x.$$

Typically, information about the regression is given by computer programs in two tables:

Table of Coefficients

Predictor	$\hat{\beta}_k$	$\mathrm{SE}(\hat{\beta}_k)$	t	P
Constant	−0.1366	0.7676	−0.18	0.860
x	0.013226	0.002263	5.84	0.000

Analysis of Variance

Source	df	SS	MS	F	p
Regression	1	26.386	26.386	34.14	0.000
Error	30	23.184	0.773		
Total	31	49.570			

Historically, the place for the Mean Squared Total is left blank but its value would be the sample variance of all n (here 32) observations in the data, ignoring any and all structure in the data.

The first table gives the estimated regression coefficients $\hat{\beta}_0 = -0.1366$ and $\hat{\beta}_1 = 0.013226$ in the second column. The first column simply identifies the terms. The third column gives a measure of variability in the estimate called the standard error. Roughly, we will be 95% confident that the

true value of the parameters is within two standard errors of the estimate. In other words, we are roughly 95% confident that β_1 will be somewhere between of $-0.1366-2(0.7676)$ and $-0.1366+2(0.7676)$. The multiplier 2 is what is "rough" about this interval. The multiplier should depend on the sample size, getting larger for small samples (giving bigger intervals and less precise knowledge) and smaller for large samples. The column t provides tests of whether the regression coefficients β_0 and β_1 are different from 0. The values are computed as $t = \hat{\beta}_k/\mathrm{SE}(\hat{\beta}_k)$. Roughly, if the value of t corresponding to β_1 is larger than 2, there is statistical evidence that $\beta_1 \neq 0$. In simple linear regression, a value of $\beta_1 = 0$, indicates that there is no relationship between x and y, i.e., the model becomes

$$y_i = \beta_0 + \varepsilon_i$$

which does not involve x. Again, the comparison value of 2 is only rough. It should change with the samples size. The final column is labeled P and gives information about the t tests. If the parameter being tested really does equal 0, P gives the probability of seeing a value of t that is as weird or weirder than what we actually saw. In this context, weird values are those that are far from 0. After all, $\hat{\beta}_k$ is an estimate of β_k, so if $\beta_k = 0$, then $\hat{\beta}_k$ should be close to 0, relative to its intrinsic variability which is measured by $\mathrm{SE}(\hat{\beta}_1)$. For testing $\beta_1 = 0$, having $P = 0.000$ means that if β_1 really is 0, we are seeing something extremely rare. There is almost no chance ($P = 0.000$) of getting a value of t as large as 5.84 when β_1 is 0. If there is a very small chance of seeing a t value this large when $\beta_1 = 0$, it follows logically that β_1 is probably not 0. Confidence intervals and tests are discussed in more detail in the next section.

The second table is called the Analysis of Variance (ANOVA) table. The entry in row "Error" and column "MS" is the mean squared error. The columns are Sources, degrees of freedom (df), sums of squares (SS), mean squares (MS), F, and P. Notice that if we add the degrees of freedom for Regression and Error we get the degrees of freedom for Total. The same thing works for the sums of squares but not the mean squares. The mean squares are just the sums of squares divided by the degrees of freedom. The F statistic provides at test of whether $\beta_1 = 0$ with values much larger than 1 indicating that $\beta_1 \neq 0$. Roughly, in this context, by much larger than 1 we mean larger than 4. Note that

$$F = 34.14 = (5.84)^2$$

where 5.84 is the t value corresponding to β_1. The P value is the same as for the t test of $\beta_1 = 0$. In this context, it is the probability of getting a value of F as large or larger than 34.14 when $\beta_1 = 0$.

An internal measure of predictive ability of a model is the coefficient of determination

$$R^2 = 53.2\% = 0.532.$$

From the SS column of the Analysis of Variance Table, $R^2 = 26.386/49.570$. It is the percentage of the total variability that is being explained by the simple linear regression model. This is an internal measure of the predictive ability in that it is based entirely on how the model fits this one set of data. An external measure of predictive ability would examine how well this model predicts for a different set of data that were collected similarly. R^2 values close to 1 indicate very strong predictive abilities and R^2 values close to 0 indicate weak predictive abilities. Note that R^2 *is not* a measure of whether this model is correct or incorrect, it is a measure of how useful the model is for making predictions. Note also that if $\beta_1 \neq 0$, the model has *some* usefulness for making predictions. R^2 is a measure of how much usefulness it has.

It should be noted that in cases like this with only one predictor variable, R^2 is the same as the square of the correlation coefficient, i.e., $R^2 = r^2$.

5.2.2 Residual Analysis

Finally, the model involves assumptions that the ε_is are independent distributed $N(0,\sigma^2)$. It is important to check these assumptions and we do so by using the residuals

$$\hat{\varepsilon}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

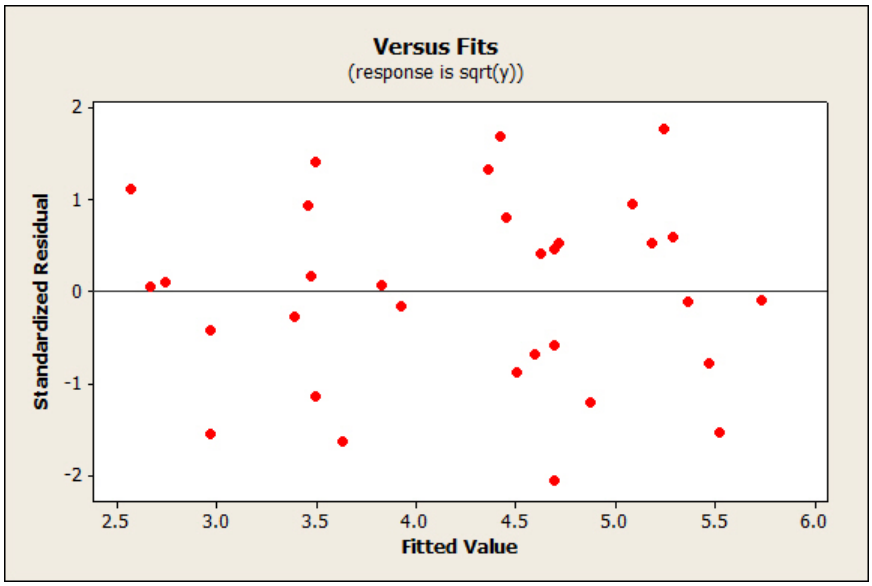

Figure 5.5: *Residuals versus predicted values, data: root y*

Actually, we most often use standardized residuals

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE(1-h_i)}}$$

where the term in the denominator is an estimate of the standard deviation of the residual, so is designed to make the standardized residual have a variance of about 1. The *leverage* h_i measures how strange a particular value of x_i is relative to the other x_is.

The residuals are often plotted against the *predicted values* (*fitted values*)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

In this plot we are looking to see an amorphous blob of points. Any systematic structure such as a curve or a horn shape indicates problems with the model. Figure 5.5 gives this plot and it seems alright.

The residuals are also often plotted against normal scores (rankits). In this plot we are looking for something that is roughly a straight line. Strong deviations from a straight line indicate problems with the normality assumption. The value of R^2 between the residuals and the normal scores is called W, to distinguish it from the R^2 between for x and y. Small values of W indicate problems with normality. $W = 0.988$ for these data, which is good. The normal scores plot is given as Figure 5.6. A little of the theory behind normal plots is presented in Section 11.4 (in order to motivate methods for analyzing data that do not provide a MSE).

5.2.3 Prediction

The prediction of $\sqrt{y}$ at 325 degrees is

$$4.162 = -0.137 + 0.0132(325).$$

Transforming back to the original scale gives a prediction of $4.162^2 = 17.322\%$. A 95% prediction interval is (2.338, 5.985) which transforms back to (5.47, 35.82). In other words, we are 95%

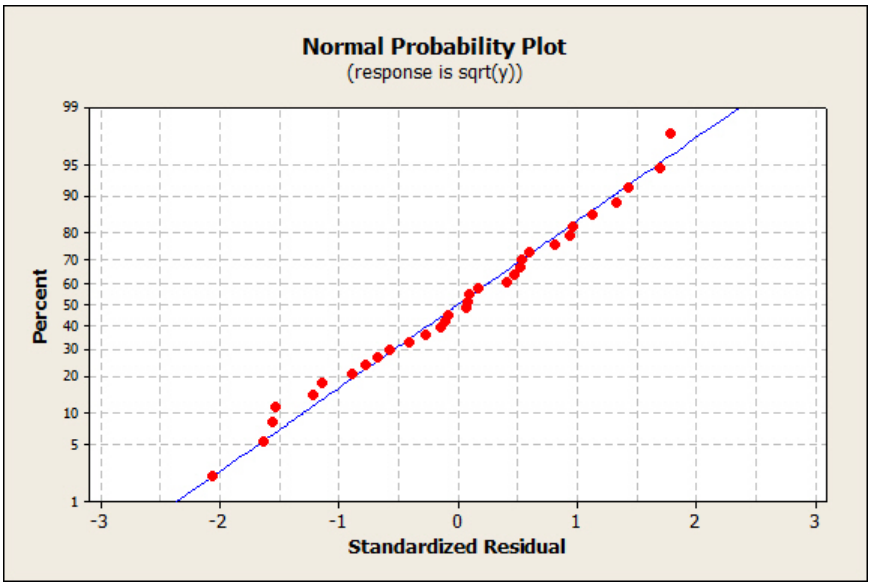

Figure 5.6: *Normal Plot for root y; W* = 0.988

confident that a new observation at 325 degrees would give a gasoline yield of between 5.47% and 35.82%. In an effort to get something in line with the crude intervals of section 1, an 80% prediction interval is (2.99225,5.33175) which transforms back to (8.95, 28.42), not too far from the values of Section 1.

Again, the 95% prediction intervals are roughly the prediction plus and minus 2 times an appropriate standard error. The prediction is an estimate of the point on the line $\beta_0 + \beta_1(325)$. There are actually two different issues of interest relative to this value. One is just to estimate the value and to measure how much variability there is in the estimate. The other is to use the estimate as a prediction for a new observation at 325 degrees and to measure how much variability there is for predicting a new observation. Note that when used for prediction, the variability will be greater because there is variability associated with using the estimate of the value of the line at 325 degrees plus there is variability that is intrinsic to taking a new observation.

For these data the standard error for the value of the line at 325 is SE(line) = 0.156. A 95% confidence interval for $\beta_0 + \beta_1(325)$ is (3.843,4.481). Note that this is roughly $4.162 \pm 2(0.156)$. Actually, it is roughly $4.162 \pm 2.04(0.156)$. On the original scale this transforms back to $(3.843^2, 4.481^2)$ or (14.77, 20.08).

To get a prediction interval we need

$$\text{SE(pred.)} = \sqrt{MSE + \text{SE(line)}^2} = \sqrt{0.773 + (0.156)^2} = 0.8929.$$

so the 95% prediction interval is roughly $4.162 \pm 2(0.8929)$, or is precisely $4.162 \pm 2.04(0.8929)$, which is (2.338, 5.985). The 80% prediction interval involves changing the multiplier from around 2 to 1.31. In particular, the interval is $4.162 \pm 1.31(0.8929)$.

It turns out that for predicting at a point x,

$$\text{SE(line)}^2 = MSE\left[\frac{1}{n} + \frac{(x - \bar{x}_\cdot)^2}{(n-1)s_{x^2}}\right].$$

5.3 Statistical Tests and Confidence Intervals

In statistics, we frequently establish *models* to help explain the behavior of data. Often we want to check whether those models seem reasonable. The purpose of a statistical test is to check whether the data are consistent with a specified model. The model establishes certain predictions about how data should behave. We then collect data and check whether the data are consistent with the predictions. In particular, many statistical tests are based on identifying a parameter of interest (*Par*), an estimate (*Est*) of that parameter, a measure of the variability of the estimate [SE(*Est*)], and a distribution for $(Est - Par)/\mathrm{SE}(Est)$. For example, with the data of Example 5.2.1, suppose we hypothesize a simple linear regression model in which the slope parameter β_1 equals 0.02. We want to check whether the data are consistent with this hypothesized model, or whether the data (tend to) contradict it. If our interest focuses on β_1, we have a parameter of interest ($Par \equiv \beta_1$). We already have an estimate of β_1, $\hat{\beta}_1 = 0.013226$. If our model with $\beta_1 = 0.02$ is correct, the estimate $\hat{\beta}_1$ should be close to 0.02. We evaluate this by checking whether $0.013226 - 0.02$ is close to zero. Is $0.013226 - 0.02$ is close to zero? That actually depends on the variability associated with the estimate of $\hat{\beta}_1$. If our model is correct, in repeated sampling $(\hat{\beta}_1 - 0.02)/\mathrm{SE}(\hat{\beta}_1)$ should have a $t(dfE)$ distribution, where $dfE = 30$ in Example 5.2.1. This is the behavior predicted by the model. We check this against the data actually collected. We actually have collected data that give us $\hat{\beta}_1 = 0.013226$ and $\mathrm{SE}(\hat{\beta}_1) = 0.002263$, so if our model is correct $(0.013226 - 0.02)/0.002263 = -2.99$ should be one observation from a $t(30)$ distribution. In sampling from a $t(30)$, 99.44% of the time we would get an observation closer to zero than -2.99. Only 0.56% of the time would we get something farther from zero than -2.99. The value -2.99 does not seem like it could reasonably come from a $t(30)$ distribution, as it was supposed to if our model were correct. The data seem to contradict the model. The number $0.56\% = 0.0056$ given above is called the P value, and is used as a measure of how strange these data are relative to the behavior predicted by the model.

Sometimes tests are presented as formal decision rules as to whether the model should be rejected or not. In such cases, an α level is chosen (usually 0.05 or 0.01), and the model is rejected if $P < \alpha$. Another way to write this procedure is to find $t(1-\alpha/2, dfE)$, the number that cuts off the top $\alpha/2$ of a $t(dfE)$ distribution. For $\alpha = 0.05$ and $dfE = 30$, this is $t(0.975, 30) = 2.0423$. The formal decision rule is to reject the model if

$$\frac{Est - 0.02}{\mathrm{SE}(Est)} = \frac{\hat{\beta}_1 - .02}{\mathrm{SE}(\hat{\beta}_1)} > 2.0423$$

of if

$$\frac{Est - 0.02}{\mathrm{SE}(Est)} = \frac{\hat{\beta}_1 - .02}{\mathrm{SE}(\hat{\beta}_1)} < -2.0423.$$

Note that the decision rule can be established without knowing the data — but using the rule requires knowing actual values for $\hat{\beta}_1$ and $\mathrm{SE}(\hat{\beta}_1)$.

This test has been focused on a particular parameter of the model, β_1. While the test is really a test of whether the entire model fits, it is particularly good at detecting differences $\beta_1 \neq 0.02$. In fact, the test is often presented as a test of a *null hypothesis* $H_0 : \beta_1 = 0.02$ versus an *alternative hypothesis* $H_A : \beta_1 \neq 0.02$. But this formulation ignores the sensitivity of the test to all of the aspects of the model other than $\beta_1 = 0.02$.

A 95% confidence interval for β_1 contains all of the parameter values β_1 that are consistent with the data (and the model) as evaluated by an $\alpha = 0.05$ test. Here, $95\% = 0.95 = 1 - .05 = 1 - \alpha$. The 95% confidence interval is every parameter value β_1 that cannot be rejected in an $\alpha = 0.05$ test. A particular value β_1 will not be rejected if

$$-t(1-\alpha/2, dfE) < \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} < t(1-\alpha/2, dfE).$$

Table 5.2: *Prater's gasoline–crude oil data*

y	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4
6.9	38.4	6.1	220	235	24.8	32.2	5.2	236	360
14.4	40.3	4.8	231	307	26.0	38.4	6.1	220	365
7.4	40.0	6.1	217	212	34.9	40.3	4.8	231	395
8.5	31.8	0.2	316	365	18.2	40.0	6.1	217	272
8.0	40.8	3.5	210	218	23.2	32.2	2.4	284	424
2.8	41.3	1.8	267	235	18.0	31.8	0.2	316	428
5.0	38.1	1.2	274	285	13.1	40.8	3.5	210	273
12.2	50.8	8.6	190	205	16.1	41.3	1.8	267	358
10.0	32.2	5.2	236	267	32.1	38.1	1.2	274	444
15.2	38.4	6.1	220	300	34.7	50.8	8.6	190	345
26.8	40.3	4.8	231	367	31.7	32.2	5.2	236	402
14.0	32.2	2.4	284	351	33.6	38.4	6.1	220	410
14.7	31.8	0.2	316	379	30.4	40.0	6.1	217	340
6.4	41.3	1.8	267	275	26.6	40.8	3.5	210	347
17.6	38.1	1.2	274	365	27.8	41.3	1.8	267	416
22.3	50.8	8.6	190	275	45.7	50.8	8.6	190	407

Some algebra establishes that β_1 will satisfy these conditions if and only if

$$\hat{\beta}_1 - t(1-\alpha/2, dfE)\,\mathrm{SE}(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t(1-\alpha/2, dfE)\,\mathrm{SE}(\hat{\beta}_1).$$

The endpoints of the interval are

$$\hat{\beta}_1 \pm t(1-\alpha/2, dfE)\,\mathrm{SE}(\hat{\beta}_1).$$

For our data with $dfE = 30$ and $\alpha = 0.05$, this is $0.013266 \pm 2.0423(0.002263)$ for an interval (0.00864, 0.01789).

In more generality, if you have a *Par* of interest, with *Est* and $\mathrm{SE}(Est)$, and the distribution $(Est - Par)/\mathrm{SE}(Est) \sim t(dfE)$, then a, say, 99% confidence interval will have limits

$$Est \pm t(1-\alpha/2, dfE)\,\mathrm{SE}(Est),$$

where $\alpha = 1 - 0.99$.

5.4 Scatterplot Matrix

It is relatively easy to see the relationship between two variables simply by plotting them. With a sophisticated plotting device, one can even examine the relationships between three variables, however with more than three variables, simultaneous plotting becomes impossible. One commonly used method to overcome this is the scatterplot matrix. It is simply a matrix of all of the pairwise plots that can be constructed from a collection of variables,

EXAMPLE 5.4.1. Hader and Grandage (1958), Atkinson (1985), and Christensen (1996) have presented Prater's data on gasoline. The variables are y, the percentage of gasoline obtained from crude oil; x_1, the crude oil specific gravity $^{\circ}$API; x_2, crude oil vapor pressure measured in lbs/in^2; x_3, the temperature in $^{\circ}$F at which 10% of the crude oil is vaporized; and x_4, the temperature in $^{\circ}$F at which all of the crude oil is vaporized. The data are given in Table 5.2. Ideally, a scatterplot matrix would be a 5×5 array of plots but that it too big to fit well on a page. A scatterplot matrix of the predictor variables x_j is given in Figure 5.7 and Figure 5.8 contains plots of y versus the predictor variables.

In Section 1 we examined the simple linear regression between $\sqrt{y}$ and what we are now calling x_4. $\sqrt{y}$ was used in that example because if you look at a plot of y on x_4, (or a plot of the residuals

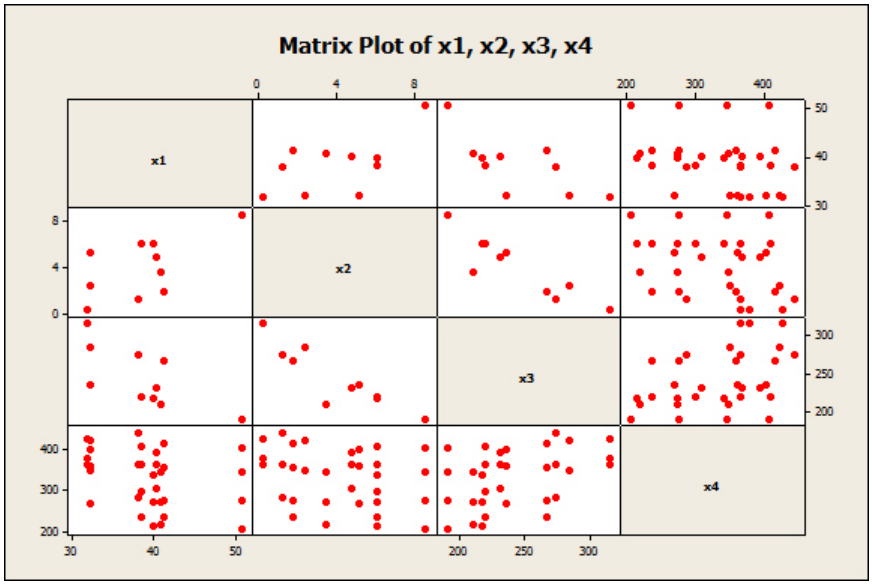

Figure 5.7: *Scatterplot Matrix for Gasoline Data: Predictor variables.*

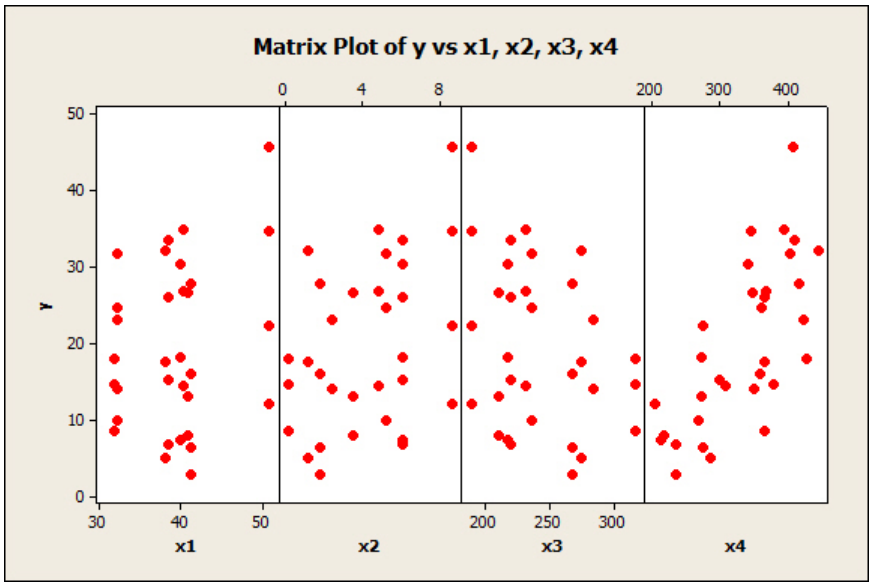

Figure 5.8: *Scatterplot Matrix for Gasoline Data: Predictor variables with y.*

versus x_4), the variability of y seems to increase with the value of x_4. Note that this plot also occurs on the right side of Figure 5.8. A log transformation did not seem to eliminate the problem, but the square root seemed to work fine. Alas, the situation is somewhat more complex. If you look at the plot of x_3 versus x_4, you see a rough linear trend in which the x_3 values show increased variability as x_4 increases. However, x_3 is also an important predictor of y that we were ignoring in Section 1. In the simple linear regression of y on x_4, the increasing variability is really a lack of fit. By ignoring the important variable x_3, we expect to see more variability. If the x_3 values are more spread out for large x_4 values than for small x_4 values, we should fit the line worse at large x_4 values than at small ones. That is exactly what we saw. It turns out that if you do a regression of y on both x_3 and x_4, you do not see the same serious problem with residual variability.

5.5 Multiple Regression

We now want to extend the ideas of simple linear regression to include all 4 of the predictor variables included in the Prater data. A multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \quad i = 1, \ldots, 32\,.$$

In general, with n observations and p predictor variables the model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n\,, \tag{5.5.1}$$

where ε_is are assumed to be independent $N(0, \sigma^2)$.

EXAMPLE 5.5.1. For the Prater data we examine a regression of y on x_1, x_2, x_3, and x_4:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i\,. \tag{5.5.2}$$

The estimated regression equation is

$$\hat{y} = -6.8 + 0.227 x_1 + 0.554 x_2 - 0.150 x_3 + 0.155 x_4$$

with tables of statistics

Table of Coefficients

Predictor	*Est*	SE	*t*	*P*
Constant	−6.82	10.12	−0.67	0.506
x_1	0.22725	0.09994	2.27	0.031
x_2	0.5537	0.3698	1.50	0.146
x_3	−0.14954	0.02923	−5.12	0.000
x_4	0.154650	0.006446	23.99	0.000

Analysis of Variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	4	3429.27	857.32	171.71	0.000
Error	27	134.80	4.99		
Total	31	3564.08			

In the coefficients table, we have the least squares (maxium likelihood) estimates of each parameter, the standard error of each parameter, the t statistic for testing whether the coefficient is equal to 0, and the P value for the test. Note that for x_3 the t statistic is -5.12 which is far from 0 and very inconsistent with β_3 being equal to 0, as measured by the very small P value. Thus it is clear that x_3 is needed to help explain the data. On the other hand, there is no strong evidence that $\beta_2 \neq 0$ and the evidence that $\beta_1 \neq 0$ is much weaker than for β_3 and β_4. So there is no strong evidence that x_2 is needed in the model and relatively weak evidence that x_1 is needed. A key fact about this table is

that everything depends on everything else. For example, if we decided that a variable should only be in the model if the corresponding P value is less than 0.01, we could conclude that *either* x_1 *or* x_2 can be dropped from the model. *But we cannot conclude that both should be dropped from the model*. The test for whether $\beta_2 = 0$ is based on having x_1 in the model and similarly the test for whether $\beta_1 = 0$ is based on having x_2 in the model. If you want to find out whether you can drop both x_1 and x_2 you have to fit a model that drops both variables and compare it to a model that has all 4 variables. A method for making such comparisons will be discussed later.

In the ANOVA table, the regression has 4 degrees of freedom because there are 4 predictor variables. The degrees of freedom for error is $27 = 32 - 5$ where the $n = 32$ is the number of observations in the data and $p + 1 = 5$ is the number of β_k coefficients being fitted in the model. The degrees of freedom total is $n - 1 = 32 - 1 = 31$ and the sum of squares total is 31 times the sample variance of the 32 y values. The sum of squares error is the sum of squared residuals. The residuals are

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \hat{\beta}_4 x_{i4}$$

and

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \hat{\beta}_4 x_{i4})^2.$$

The mean squared error is

$$MSE = SSE/dfE = 134.80/27 = 4.99.$$

The sum of squares regression is the sum of squares total minus the sum of squares error,

$$SSReg = 3564.08 - 134.80 = 3429.27$$

The F statistic provides a test of whether any of the predictor variables is useful in explaining the data, i.e., it tests whether we could drop all of x_1, x_2, x_3, and x_4 and still do an adequate job of explaining the data, i.e., it tests whether $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. The test is rejected if the F value is much larger than 1. In particular, the test is rejected if, assuming the variables are *not* needed, the probability of seeing an F value as large or larger than the one we actually saw is too small, i.e., if the P value is too close to 0.

The R^2 for these data is

$$R^2 = \frac{SSReg}{SSTot} = \frac{3429.27}{3564.08} = 96.2\%.$$

Consider a prediction for the values $x_1 = 44$, $x_2 = 6$, $x_3 = 230$, $x_4 = 230$. Using the estimated regression equation, the point prediction is

$$7.677 = -6.8 + 0.227(44) + 0.554(6) - 0.150(230) + 0.155(230).$$

Typical computer output provides a standard error for the fitted surface, in this example it is $\mathrm{SE}(surf) = 0.953$. As discussed in Section 2, this standard error must be modified to give a standard error appropriate for a prediction interval,

$$\mathrm{SE}(pred) = \sqrt{MSE + [\mathrm{SE}(surf)]^2} = \sqrt{4.99 + [0.953]^2)} = 2.429$$

The 95% prediction interval is approximately 7.677 plus or minus 2 times this standard error. Here the exact interval involves a multiplier slightly larger than 2, so the exact interval is $(2.692, 12.662)$, cf. Christensen (1996, Secs. 13.2 and 15.4).

The sequential sums of squares are

Source	df	Seq SS
x1	1	216.26
x2	1	309.85
x3	1	29.21
x4	1	2873.95

5.6 Polynomial Regression

We can fit more general models than those used in the previous section. In particular we can use the same methods for fitting quadratic models. For simplicity, we focus on only two variables. For the Prater data we focus on x_3 and x_4. We now consider a quadratic model

$$y_i = \beta_{00} + \beta_{10}x_{i3} + \beta_{01}x_{i4} + \beta_{20}x_{i3}^2 + \beta_{02}x_{i4}^2 + \beta_{11}x_{i3}x_{i4} + \varepsilon_i\,.$$

The fitted regression equation is

$$\hat{y} = 40.6 - 0.371x_3 + 0.133x_4 + 0.000687x_3^2 + 0.000222x_4^2 - 0.000519x_3x_4.$$

The tables of statistics are similar to those used earlier.

Table of Coefficients

Predictor	*Est*	SE	*t*	*P*
Constant	40.59	19.78	2.05	0.050
x_3	−0.3712	0.1462	−2.54	0.017
x_4	0.13330	0.06214	2.15	0.041
x_3^2	0.0006872	0.0003288	2.09	0.047
x_4^2	0.0002224	0.0001114	2.00	0.056
x_3x_4	−0.0005187	0.0002380	−2.18	0.039

Analysis of Variance

Source	*df*	*SS*	*MS*	*F*	*P*
Regression	5	3431.38	686.28	134.46	0.000
Error	26	132.70	5.10		
Total	31	3564.08			

$$R^2 = 96.3\%$$

In addition, one should examine residual plots and one can examine predictions.

Figures 5.9 and 5.10 plot the regression surface. In experimental design such polynomial surfaces are referred to as *response surfaces* and are the subject of considerable theory, cf *TiD*.

5.7 Matrices

Computations and theory for regression are greatly simplified by using matrices. Virtually all computer programs for doing regression are matrix oriented. We briefly introduce matrix notation for regression problems. *This section presumes that the reader knows how to add and multiply matrices. Near the end we also use the concepts of the transpose of a matrix and of the inverse of a square matrix.* A vector is a matrix with just one column.

EXAMPLE 5.7.2. *Simple Linear Regression.*
Consider the relationship between y, the percent of gasoline obtained from crude oil, and x, the temperature in °F at which all the crude oil is vaporized. Five observations are given below.

x_i	y_i
307	14.4
365	26.0
235	2.8
428	18.0
345	34.7

We can fit the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,\ldots,5,$$

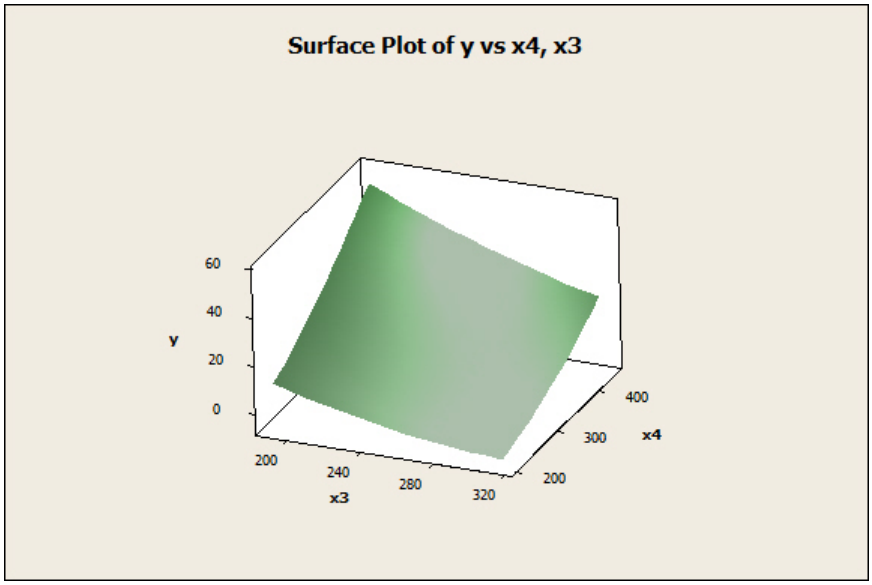

Figure 5.9: *Surface plot of y.*

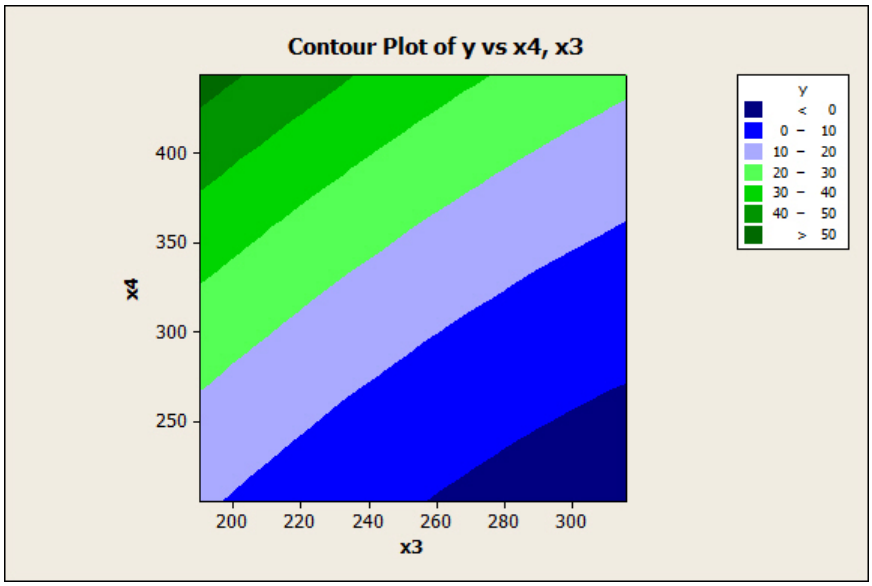

Figure 5.10: *Contour plot of y.*

where the ε_i's are assumed to be independent $N(0,\sigma^2)$. In matrix notation we can write the model as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} \beta_0+\beta_1(307)+\varepsilon_1 \\ \beta_0+\beta_1(365)+\varepsilon_2 \\ \beta_0+\beta_1(236)+\varepsilon_3 \\ \beta_0+\beta_1(428)+\varepsilon_4 \\ \beta_0+\beta_1(345)+\varepsilon_5 \end{bmatrix}. \tag{5.7.1}$$

These two matrices are equal if and only if the corresponding elements are equal, which occurs if and only if the simple linear regression model holds. The x_i's are assumed to be fixed known numbers, so we have incorporated them into the statement of the model. The y_i's are assumed to be random variables, the values given earlier for the y_i's are realizations of the random variables. We use them to fit the model (estimate parameters), but they are not part of the model itself.

We can also write the simple linear regression model as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 307 \\ 1 & 365 \\ 1 & 235 \\ 1 & 428 \\ 1 & 345 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$Y = X \beta + e$$

Multiplying and adding the matrices on the right-hand side establishes that this is equivalent to equation (5.7.1). Note that we observe the Y vector to be

$$Y = (14.4, 26.0, 2.8, 18.0, 34.7)'.$$

In general, the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$, can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0+\beta_1 x_1+\varepsilon_1 \\ \beta_0+\beta_1 x_2+\varepsilon_2 \\ \vdots \\ \beta_0+\beta_1 x_n+\varepsilon_n \end{bmatrix}$$

or, equivalently,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times 2} \beta_{2\times 1} + e_{n\times 1}$$

The usual conditions are that the ε_i's are independent $N(0,\sigma^2)$. We can restate this as that (a) all ε_is are independent, (b) $\mathrm{E}(\varepsilon_i) = 0$ for all i, (c) $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i, (d) all are normally distributed. Condition (b) translates into matrix terms as

$$\mathrm{E}(e) \equiv \begin{bmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0_{n\times 1},$$

where 0 is the $n \times 1$ matrix containing all zeros. Conditions (a) and (c) imply that

$$\mathrm{Cov}(e) = \sigma^2 I,$$

where I is the $n \times n$ identity matrix. Here $\mathrm{Cov}(e)$ is defined as the $n \times n$ matrix with ij element $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$. By definition, the covariance matrix $\mathrm{Cov}(e)$ has the variances of the ε_i's down the diagonal. The variance of each individual ε_i is σ^2, so all the diagonal elements of $\mathrm{Cov}(e)$ are σ^2, just as in $\sigma^2 I$. Also by definition, the covariance matrix $\mathrm{Cov}(e)$ has the covariances of distinct ε_i's as its off-diagonal elements. The covariances of distinct ε_i's are all 0 because they are independent, so all the off-diagonal elements of $\mathrm{Cov}(e)$ are zero, just as in $\sigma^2 I$.

In general linear model is any model

$$Y = X\beta + e,$$

where Y is an observable random vector, X is fixed and known, β is fixed but unknown, so e is the only thing random on the right of the equation and $\mathrm{E}(e) = 0$. It follows that

$$\mathrm{E}(Y) = \mathrm{E}(X\beta + e) = X\beta + \mathrm{E}(e) = X\beta$$

and, with our usual assumptions about the ε_is,

$$\mathrm{Cov}(Y) = \mathrm{Cov}(X\beta + e) = \mathrm{Cov}(e) = \sigma^2 I.$$

These results follow directly from properties enumerated in Subsection 3.3.3 but, more to the point, they are intuitive since the vector $X\beta$ is fixed and has no variability. $\mathrm{Cov}(e) = \sigma^2 I$ is a common but not necessary assumption for having a linear model.

EXAMPLE 5.7.3. *Polynomial Regression.*
Rather than fitting a line to the (x, y) data of Example 5.7.2, we can fit, say, a quadratic model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \qquad i = 1, \ldots, 5,$$

where the ε_i's are assumed to be independent $N(0, \sigma^2)$. In matrix notation we can write the model as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1(307) + \beta_2(307^2) + \varepsilon_1 \\ \beta_0 + \beta_1(365) + \beta_2(365^2) + \varepsilon_2 \\ \beta_0 + \beta_1(236) + \beta_2(236^2) + \varepsilon_3 \\ \beta_0 + \beta_1(428) + \beta_2(428^2) + \varepsilon_4 \\ \beta_0 + \beta_1(345) + \beta_2(345^2) + \varepsilon_5 \end{bmatrix}.$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 307 & 307^2 \\ 1 & 365 & 365^2 \\ 1 & 235 & 236^2 \\ 1 & 428 & 428^2 \\ 1 & 345 & 345^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$Y = X \beta + e.$$

In general, the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $i = 1, \ldots, n$ can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \varepsilon_n \end{bmatrix}$$

or, equivalently, as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Similarly, the cubic model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$, $i = 1, \ldots, n$, can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \beta_3 x_n^3 + \varepsilon_n \end{bmatrix}$$

or, equivalently, as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times 4} \beta_{4\times 1} + e_{n\times 1}$$

Other polynomials follow the same pattern.

EXAMPLE 5.7.5. *Multiple Regression.*
Consider the relationship between y, the percent of gasoline obtained from crude oil, x_1, the temperature in °F at which 10% of the crude oil is vaporized, and x_2, the temperature at which all the crude oil is vaporized. Five observations are given below.

x_{i1}	x_{i2}	y_i
231	307	14.4
220	365	26.0
267	235	2.8
316	428	18.0
190	345	34.7

The multiple regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1,2,3,4,5.$$

In matrix terms the model can be rewritten as the equality of two 5×1 matrices,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1(231) + \beta_2(307) + \varepsilon_1 \\ \beta_0 + \beta_1(220) + \beta_2(365) + \varepsilon_2 \\ \beta_0 + \beta_1(267) + \beta_2(235) + \varepsilon_3 \\ \beta_0 + \beta_1(316) + \beta_2(428) + \varepsilon_4 \\ \beta_0 + \beta_1(190) + \beta_2(345) + \varepsilon_5 \end{bmatrix}.$$

Breaking up the right-hand-side gives

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 231 & 307 \\ 1 & 220 & 365 \\ 1 & 267 & 235 \\ 1 & 316 & 428 \\ 1 & 190 & 345 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times p} \beta_{p\times 1} + e_{n\times 1}.$$

We could also fit a polynomial in these two variables, say,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i.$$

In matrix form this is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 231 & 307 & 231^2 & 307^2 & 231(307) \\ 1 & 220 & 365 & 220^2 & 365^2 & 220(365) \\ 1 & 267 & 235 & 267^2 & 235^2 & 267(235) \\ 1 & 316 & 428 & 316^2 & 428^2 & 316(428) \\ 1 & 190 & 345 & 190^2 & 345^2 & 190(345) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}.$$

The general multiple regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

In matrix terms the model involves equality of two $n \times 1$ matrices,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{p-1} x_{1,p-1} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_{p-1} x_{2,p-1} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{p-1} x_{n,p-1} + \varepsilon_n \end{bmatrix}.$$

Breaking up the right-hand-side gives

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times p} \beta_{p\times 1} + e_{n\times 1}$$

Multiplying and adding the right-hand side gives the equivalence.

Once again, the usual conditions on the ε_i's translate into

$$\mathrm{E}(e) = 0,$$

where 0 is the $n \times 1$ matrix consisting of all zeros, and

$$\mathrm{Cov}(e) = \sigma^2 I,$$

where I is the $n \times n$ identity matrix.

For a general linear model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I,$$

it is not too difficult to show that the least squares estimates satisfy

$$\hat{\beta} = (X'X)^{-1} X'Y,$$

where X' denotes the transpose of the matrix X and, when such a matrix exists, $(X'X)^{-1}$ is the unique matrix that satisfies $(X'X)^{-1}(X'X) = I$. The sum of squares for error is

$$SSE \equiv (Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

The variances of the $\hat{\beta}_j$s are the diagonal elements of the matrix

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

The last three displayed matrix equations are enough to determine both the Table of Coefficients and the ANOVA table.

Table 5.3: *Regression*

Output	Rate	Align	B gap	Proll	Air	Temp.	Calrd
1.83	68	2	1	0.95	2	2	2
−3.17	76	1	1	1.05	2	2	1
−3.92	68	1	2	0.95	2	1	1
0.83	76	2	1	0.95	1	2	2
−3.92	76	1	2	0.95	1	1	1
−2.08	85	1	2	1.05	1	2	2
−3.08	85	2	2	1.05	2	1	1
−4.17	68	2	2	1.05	1	2	1
1.50	76	2	2	1.05	2	1	2
−1.58	85	1	1	0.95	2	2	2
−3.83	85	2	1	0.95	1	1	1
1.50	68	1	1	1.05	1	1	2

5.8 Logistic Regression

Test for lack of fit of intercept only in any binomial problem from Chapter 4. Test for linear time trend.

Table 4.18: Haagen-Daz special blend. One-way and Two way. Poisson model?

5.9 Missing Data

Hand, David J. *Dark Data*

Missing at random. Remaining data are still a random sample.

Missing because of data you have. (x,y) plot, not allowed to collect y if x is too low. Don't complete items that are unlikely to meet specs.

Missing because of the data you would have got. (x,y) plot, items with low y not measured. Toss out items that look like they will not meet specs.

5.10 Exercises

Carter qe96-502
Rate: 68 or 76
Alignment: right 1 inch (2) or left 1 inch (1)
B gap: loose (2) or tight (1)
Proll: 0.95 or 1.05
Air: closed (2) or open (1)
Room Temp.: hot (2) or cold (1)
Calrd: on (2) or off (1)

Table 5.4: *DBP Measurements of Carbon Black*

Time (Hour)	At Smoke Header	End Product
7	94.8	90.5
9	94.8	91.1
11	94.2	89.1
13	94.4	89.2
15	93.4	89.3
17	94.2	89.2
19	94.6	87.3
21	96.6	88.2
23	96.3	89.7
1	96.6	91.1
3	96.7	90.6

Schneider and Pruett (1994). qe94-345

predict end product dbp from earlier smoke header dbp.

Plot versus lags 0, 1, 2, 3. (lag 1 means 2 hours) get decent correlation at lag 3 but looks weird in plot. relationship is all from last 3 points with much larger smoke headers.

Table 5.5: *Relation between two variables*

Case	Bending Strength	Sturdiness	Case	Bending Strength	Sturdiness
1	170	111	11	171	113
2	168	106	12	160	105
3	154	108	13	161	115
4	170	115	14	160	107
5	161	114	15	166	117
6	173	113	16	166	111
7	172	112	17	171	108
8	175	114	18	162	103
9	164	109	19	157	106
10	160	106	20	159	107

Shi and Lu(1994). qe94-526 SLR

Chapter 6

Time Series

In an industrial or service process, as discussed in Chapter 4, our ideal is that data generated by the process be independent and identically distributed. In particular, the data would all have the same mean value and variance. Processes that display these characteristics are *under control*. The object of this desire for having processes under control is that, when the data are independent and identically distributed, we can make predictions about the future of the process. If the process displays changes in the mean value, or the variance, or even in the distribution, often the causes of such changes can be identified and eliminated and the process put back under control. Alas, sometimes (perhaps often) the problem with a process is that the observations are not independent. Means charts are sensitive to this problem, i.e., they will tend to go out of control more often if the observations have positive correlation. But lack of independence is typically more difficult to rectify than are problems with changing means or variances. Fortunately, lack of independence is often a problem that one can simply live with. Rather than requiring that observations be independent, we can require that the observations be *stationary*.

A sequence of observations is called stationary if the statistical properties of the observations remain the same regardless of what time you choose to look at them. The word "stationary" is a reference to the fact that these properties do not change. If you have a stationary process, the mean values remain the same over time and the variances remain the same over time, just like they would if the observations were independent and identically distributied. Really, the only assumption that is being weakened with stationary observations is the assumption of independence. The beauty of having a stationary process is that one can still make predictions about the future, because the conditions in the future should be the same as those in the past when we collected our data. As a matter of fact, a formal definition of a stationary process includes observations that are independent and identically distributed as a special case of a stationary process. So the methods of Chapter 4 are methods for a special case of stationary processes.

We need two things to work with stationary processes. First, we need to be able to identify situations in which observations are stationary but not independent. Second, we need to have control charts for stationary processes. The next section deals with methods for detecting lack of independence.

EXERCISE 6.0.1. Watch the video of George Box rethinking quality control and be prepared to answer very simple questions about it. (You can skip the 15 minutes of Q&A.) He discusses stationary processes, EWMAs, and suggests control charting the estimated errors (residuals) of models like those discussed below. He has some attitudes that are very different from Shewhart/Deming in that at one points he essentially advocates plotting against limits that the process owner would find disturbingly off target rather than worrying about whether the process is under control. Recall that EWMA charts are really tests for being on target. Along these lines, the posted comments surrounding the video provide some perspective. Insightful people tend to do better statistics than dogmatic people, no matter how good the dogma is. (And there are some very good statistical dogmas.)

Box quite famously believes that all models are wrong but that some are useful. Not surprisingly, he believes that no real processes are stationary. That brings us back to the issue of finding an

Table 6.1: *Temperature Data*

0.64	0.94	0.91	0.92	0.89	1.04	1.07	0.99	0.95	1.05
1.14	1.36	1.15	1.25	1.33	1.39	1.35	1.41	1.04	0.93
1.05	1.53	1.88	1.80	1.65	2.05	2.33	1.72	1.48	1.66
1.28	0.93	0.87	1.05	1.18	0.86	1.16	1.21	1.22	1.69
1.59	1.69	0.93	0.57	1.03	1.54	0.98	0.84	0.78	1.03
1.05	1.31	1.46	1.56	0.91	1.16	1.02	0.80	1.38	1.70
1.88	2.24	2.03	2.00	2.32	1.59	1.03	1.46	1.93	1.57
1.20	0.82	1.39	1.32	1.52	1.08	1.16	1.41	1.09	0.89

Read across then down.

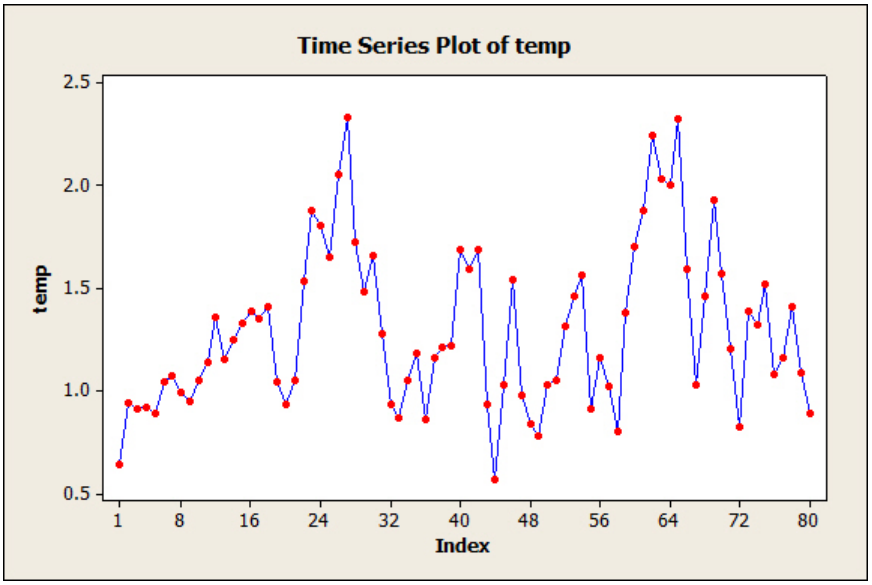

Figure 6.1: *Plot of temperature data versus time*

operational definition of when a process is close enough to being stationary that we can make useful predictions.

6.1 Autocorrelation and Autoregression

The next example illustrates a technique from time series analysis for detecting lack of independence in a series of observations.

EXAMPLE 6.1.1 The data in Table 6.1 are from Box and Lucero (1997). They consist of 80 temperatures taken "from the output of an industrial process." Figure 6.1 presents the standard time series plot of the data versus the time at which the observations were obtained. Since we are interested in industrial statistics, Figure 6.2 presents the corresponding individuals control chart. The process is wildly out of control. Not only do observations get outside the control limits but Minitab's supplemental tests 2 and 6 are violated many times and test 8 is also violated. Despite this fact, one can make a decent case for these data being the result of a stationary process and therefore predictable and capable of yielding temperatures that could be within some specifications.

Figure 6.3 plots each value against the previous measurement, i.e., it plots the pairs (y_t, y_{t-1}). There is an obvious linear relationship, leading us to expect that the current temperature should give us some ability to predict the next temperature.

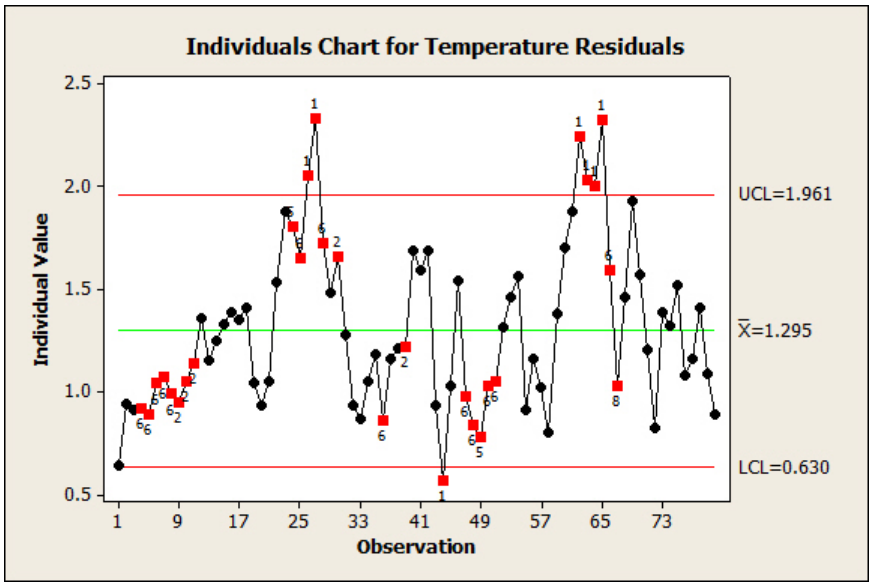

Figure 6.2: *Individuals control chart: Temperature data. (Not "residuals.")*

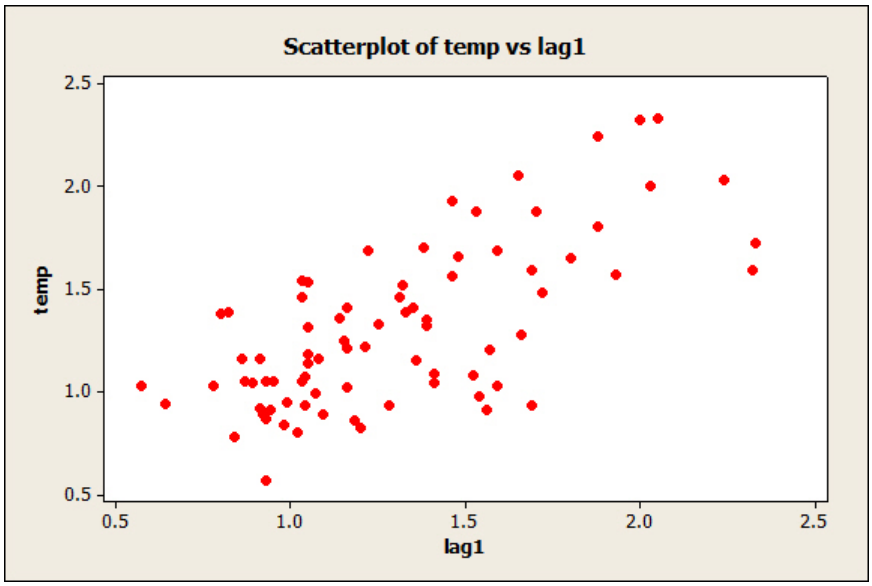

Figure 6.3: *Plot of y_t versus y_{t-1}: Temperature data.*

If there is a lag 1 linear relationship, there should be a (typically weaker) lag 2 linear relationship. Figure 6.4 plots each value against the measurement *prior to* the previous measurement, i.e., it plots the pairs (y_t, y_{t-2}). Again, there is something of a linear relationship but it is not as strong as in Figure 6.3.

The linear relationship visible is Figure 6.3 suggests that we might want to perform a regression of y_t on y_{t-1}, which in a sense, is a regression of y_t on itself, and so is called an *autoregression*. This first order autoregression model can be written as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t.$$

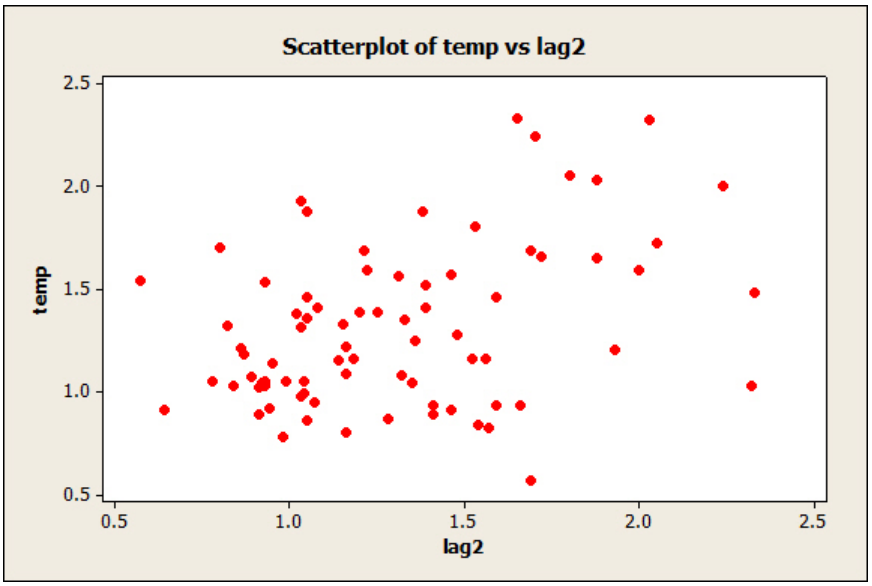

Figure 6.4: *Plot of y_t versus y_{t-2}: Temperature data.*

Often the names of the β parameters are changed to

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t.$$

Since there also seems to be a linear relationship between y_t and y_{t-2}, we might consider fitting the multiple regression model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t.$$

This is also referred to as a second order autoregressive model and rewritten as

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t.$$

Figures 6.3 and 6.4 look at the simple linear relationship between y_t and y_{t-1} and between y_t and y_{t-2} respectively. As discussed in Chapter 5, the linear relationship in these plots can be measured by something called the (sample) correlation coefficient. Figure 6.5 gives a plot of the autocorrelation function (ACF). The horizontal axis is the "lag" and the vertical axis is the correlation coefficient. At lag k it is the sample correlation between the pairs (y_t, y_{t-k}). For instance, at lag 1, the correlation is 0.664, thus the correlation for the data in Figure 6.3 is 0.664. Similarly, at lag 2, the correlation is 0.358, so the correlation for the data in Figure 6.4 is 0.358. The correlation at lag 17 is -0.262 so a plot of y_t versus y_{t-17} would display a correlation of -0.262. The graph in Figure 6.5 is simply a visual display of the (auto)correlations. (In computing autocorrelations there are fewer data pairs available for computing a lag 20 correlation than for a lag 1 correlation so corrections, or their lack, get involved in computing the ACF.)

To determine if a correlation is significantly different from zero, when the sample size n is large the estimated correlation can be compared to its approximate (null) standard error, $1/\sqrt{n}$. Thus, in this example, a correlation larger than two standard deviations, about $2/\sqrt{80} = 0.224$, is evidence for lack of independence. Figure 6.5 gives several correlations that provide evidence for a lack of independence. (The limits in Figure 6.5 are fine tuned so they get a little larger as k gets larger.)

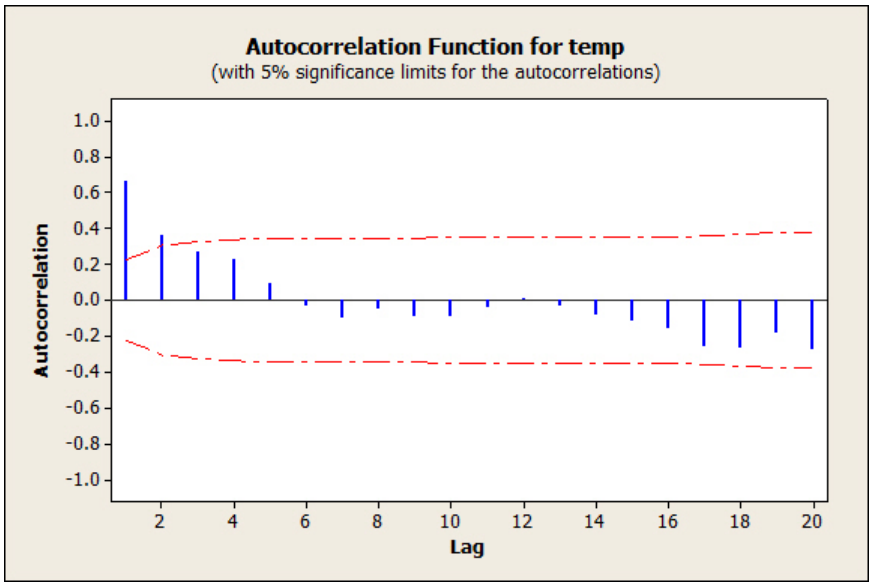

Figure 6.5: *Autocorrelation Function for Temperature Data*

6.2 Autoregression and Partial Autocorrelation

Earlier we discussed that based on Figures 6.3 and 6.4 we might consider fitting either of the regression models

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$$

or

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t.$$

The partial autocorrelation function (PACF) displayed in Figure 6.5 can help us decide between the two models. At lag 1, the value in the PACF is identical to the value in the ACF, it is just the correlation between y_t and y_{t-1}. However, the value of the PACF at lag 2 is very different from the value of the ACF at lag 2. At lag 2 the ACF is just the correlation between y_t and y_{t-2}. At lag 2 the PACF is the correlation between the residuals from fitting y_t on y_{t-1} and the residuals from fitting y_{t-2} on y_{t-1}. The idea is that if y_{t-2} really is needed in the regression model, then there should be a substantial correlation between these two sets of residuals. If there is no such correlation between the residuals, it suggests that the linear relationship we are seeing in Figure 6.4 is merely an artifact. From Figure 6.3, we know that y_t and y_{t-1} are linearly related, which immediately implies that y_{t-1} and y_{t-2} are linearly related, it follows that there must be some linear relationship between y_t and y_{t-2}. The PACF at lag 2 is looking to see if there is any direct relationship between y_t and y_{t-2} over and above the artifactual one induced by them both being related to y_{t-1}. In fact, the correlation of -0.150 is quite small, so there is no evidence that we need to include y_{t-2} as a predictor variable in our regression.

In general, the lag k partial autocorrelation is the the sample correlation between the residuals from regressing y_t on $y_{t-1}, \ldots, y_{t-k+1}$ and the residuals from regressing y_{t-k} on $y_{t-1}, \ldots, y_{t-k+1}$.

A curious behavior in the PACF for these data is the relatively large value at lag 20 of -0.325. This is indicting that if we regressed y_t on all of $y_{t-1}, \ldots, y_{t-19}$ and regressed y_{t-20} on all of $y_{t-1}, \ldots, y_{t-19}$, the residuals have a correlation of -0.325. This could either be caused by a real phenomenon or could be a mere artifact of the data. If, for example, there were 20 temperature measurements made each day, large PACF values at only lags 1 and 20 indicate that the current temperature depends directly on two things, the previous temperature and the temperature at the same

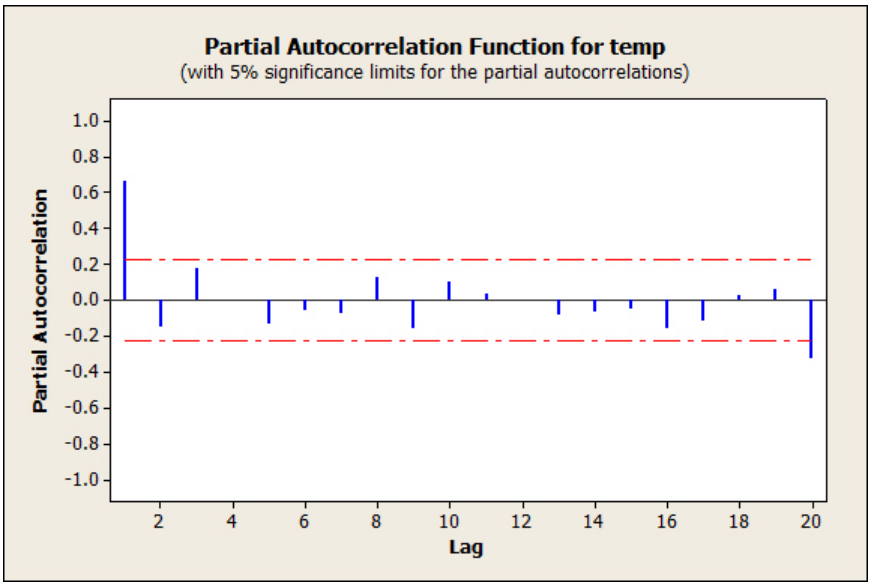

Figure 6.6: *Partial Autocorrelation Function for Temperature Data*

time the previous day. Such effects on a time series that stem from periods that are far removed are called seasonal effects. The terminology comes from the fact that many business and economic time series display regular changes associated with the seasons of the year. (Retail sales go up every Christmas, fuel oil sales for heating go down in the summer and up in the winter, etc.) We will not discuss seasonal effects further.

As indicated earlier, if you use standard regression programs to reproduce the correlations discussed here, you will probably get slightly different numbers, because the way in which correlations are computed for time series data are *slightly* different than the way they are computed in regression.

6.3 Fitting Autoregression Models

The simplest way to analyze the first order autoregressive model is to simply fit it using a standard program for regressing y_t on y_{t-1}. Note that although there are 80 observations in the series, there are only 79 pairs of (y_t, y_{t-1}) observations. The estimated regression equation is

$$\hat{y}_t = 0.428 + 0.673 y_{t-1}.$$

The usual tables of information are

Table of Coefficients

Predictor	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	t	P
Constant	0.4283	0.1121	3.82	0.000
y_{t-1}	0.67306	0.08256	8.15	0.000

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	5.5050	5.5050	66.45	0.000
Error	77	6.3786	0.0828		
Total	78	11.8836			

The coefficient of determination is $R^2 = 46.3\%$.

The predicted value for the next observation is obtained by plugging the last observation, $y_{80} = 0.89$, into the regression function:

$$\hat{y}_{81} = 0.428 + 0.673(0.89) = 1.0273.$$

The standard error for fitting the line is $\text{SE}(line) = 0.0469$, so the standard error of prediction is $\text{SE}(pred) = \sqrt{0.0828 + [0.0469]^2}$ and the 95% prediction interval is $(0.4466, 1.6080)$.

To obtain a prediction for the subsequent day, plug the prediction $\hat{y}_{81}$ into the regression equation.

$$\hat{y}_{82} = 0.428 + 0.673(1.0273) = 1.119.$$

To get a prediction interval, one would have to rerun the regression program asking for a prediction at 1.0273.

Note that if necessary, we could also have used both y_{t-1} and y_{t-2} as predictors of y_t. This would have given a multiple regression model,

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$$

with a similar analysis.

While standard regression fitting procedures work reasonably well for these autoregressive models, they are slightly inefficient and do not generalize to other useful time series models. We now give the results of an alternative model fitting algorithm that is more efficient and does generalize. This method is also based on finding parameters by using least squares, but does so in a slightly different way. This method is an iterative procedure that starts with crude estimates of the parameters and successively modifies them until the values converge to appropriate estimates, whereas the regression estimates are found in just one step. In this example, it took 8 modifications of the crude estimates before they converged. The final estimates are given in a standard table

Table of Coefficients

Type	Estimate	SE	t	P
AR 1	0.6986	0.0821	8.51	0.000
Constant	0.38165	0.03237	11.79	0.000
Mean	1.2661	0.1074		

The row for "AR 1" is for the slope estimate, and the "Constant" row is as in the regression table. Note the rough similarity of these estimates to those obtained by simply doing linear regression. The third row in this table is for the mean. It is the value that will be approached by predictions for the distant future. It is the estimate of the mean of the stationary process.

Rather than giving an Analysis of Variance table, this procedure simply reports $DFE = 78$, $SSE = 6.50357$, and $MSE = 0.08338$.

The predicted value for the next observation is obtained by plugging the last observation, $y_{80} = 0.89$, into the estimated autoregression function:

$$\hat{y}_{81} = 0.38165 + 0.6986(0.89) = 1.00337.$$

and the 95% prediction interval is given by the program as $(0.43730, 1.56945)$. Note the similarity of this point prediction and prediction interval to those obtained by doing simple linear regression.

To obtain a prediction for the subsequent day, plug the prediction $\hat{y}_{81}$ into the regression equation.

$$\hat{y}_{82} = 0.38165 + 0.6986(1.00337) = 1.08257$$

with 95% prediction interval $(0.39206, 1.77309)$.

As with any regression we examine the residuals to check if the model assumptions seem reasonable. Figures 6.7 through 6.10 are various residual plots. Figure 6.7 looks at the standard residual plots associated with regression y_t on y_{t-1}. It looks ok. The errors ε_t are assumed to be iid, so we

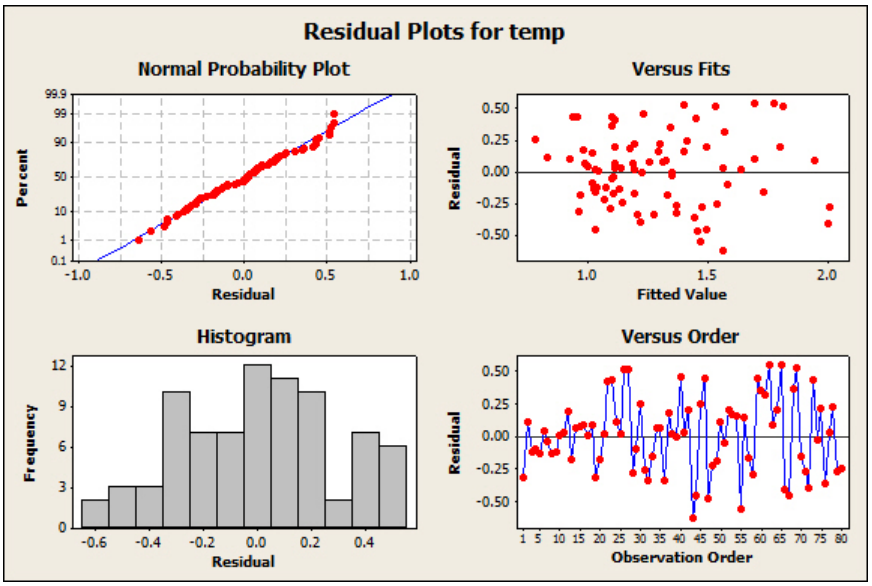

Figure 6.7: *Residual plots for Temperature Data*

can use our time series and control plots to check whether the residuals seem consistent with the errors being iid. Figures 6.8 through 6.10 give the ACF, PACF, and individuals chart applied to the residuals. They all look consistent with the errors being iid.

Finally, the program gives diagnostic tests as to how well the model fits the data. The chi-square statistics are based on a weighted sum of the squared residual autocorrelations back to, respectively, lags 12, 24, 36, and 48.

Modified Box-Pierce (Ljung-Box) χ^2 statistics				
Lag	12	24	36	48
Chi-Square	13.0	40.7	52.5	69.1
DF	10	22	34	46
P-Value	0.222	0.009	0.022	0.015

Unfortunately, these are not tremendously good for the temperature data, especially for longer lags. This probably relates to the odd behavior observed in the large partial autocorrelation at lag 20 and makes me wonder if there might be some unexpected seasonal component.

6.4 ARMA and ARIMA Models

The time series program used is designed for fitting integrated autoregressive moving average (ARIMA) models. We have only discussed autoregressive models such as the second order autoregressive model [$AR(2)$ model]

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t.$$

Typically, different parameter letters are used for fitting time series. This model is identical to

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t,$$

only the names of the parameters have been changed to confuse the reader. Moving average models are similar but involve relating current observations to the errors involved with earlier observations. A second order moving average model [$MA(2)$ model] is

$$y_t = \mu + \varepsilon_t - \psi_1 \varepsilon_{t-1} - \psi_2 \varepsilon_{t-2}.$$

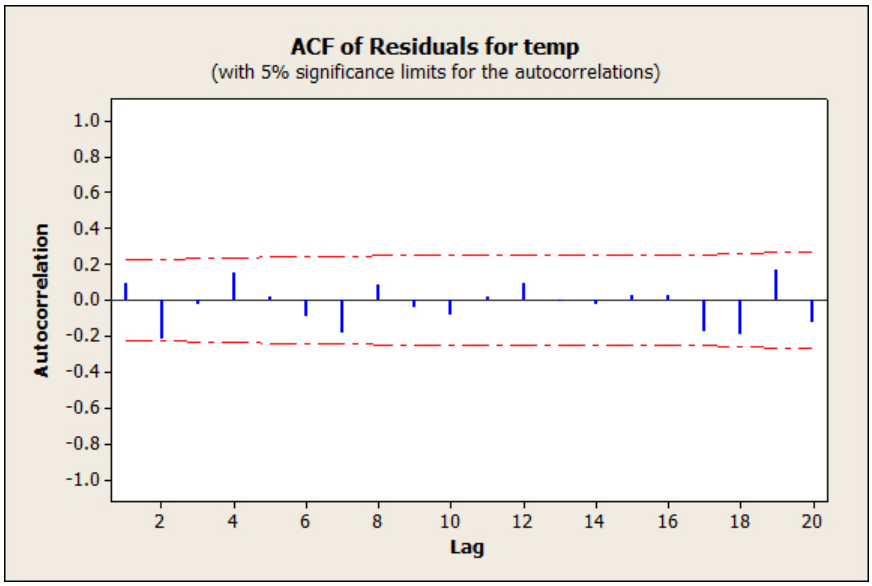

Figure 6.8: *Autocorrelation plot: Temperature Data Residuals*

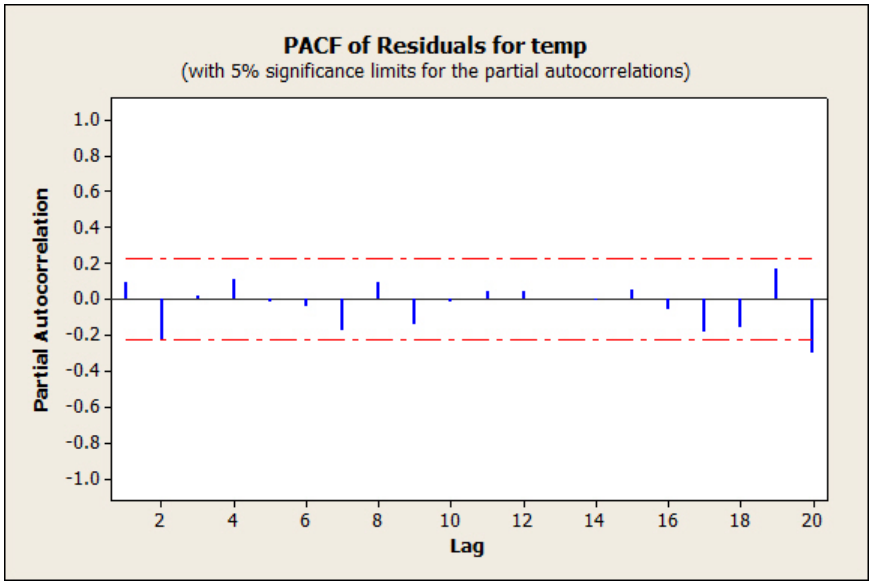

Figure 6.9: *Partial autocorrelation plot: Temperature Data Residuals*

These can be combined into an autoregressive moving average model

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t - \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2}$$

that is second order in both the AR and the MA. This is referred to as an $ARMA(2,2)$ model.

ARMA models are used to model stationary processes. To model nonstationary models, differencing is often used. For example, if a process has a linear trend over time, taking first differences of the time series, i.e., $y_t - y_{t-1}$ will eliminate the trend.

The word "integrated" in integrated autoregressive moving average models actually refers to

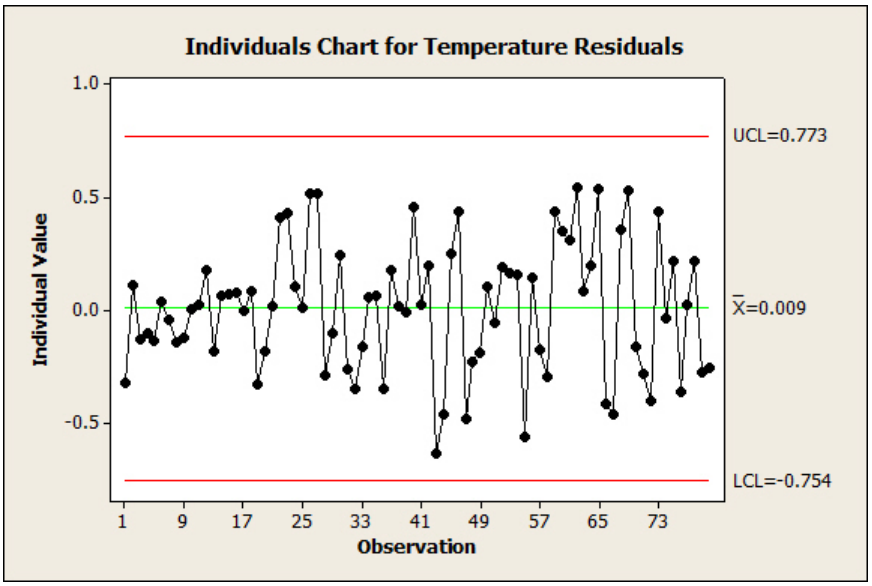

Figure 6.10: *Individuals Control Chart: Temperature Data Residuals*

differencing. An $ARIMA(2,1,2)$ model is actually an $ARMA(2,2)$ model for the series differenced series $y_t - y_{t-1}$. In general, one can fit $ARIMA(p,d,q)$ models which take d differences (a second order difference is analyzing $w_t - w_{t-1}$ where $w_t \equiv y_t - y_{t-1}$) and fit p autoregressive terms $(y_{t-1},\ldots,y_{t-p})$ and q moving average terms $(\varepsilon_{t-1},\ldots,\varepsilon_{t-q})$.

It turns out that an $ARIMA(0,0,1)$ leads to predictions that are essentially EWMAs. In the video cited earlier, Box mentions that he has found $IMA(1,1)$ ($ARIMA(0,1,1)$) models to be very useful in industrial applications.

6.5 Computing

Minitab is easy.

R code from my book *ALM-III*

```
rm(list = ls())
coal.production <- read.table(
"C:\\E-drive\\Books\\LINMOD23\\DATA\\ALM6-1.dat",
   sep="",col.names=c("Year","Wt"))
attach(coal.production)
coal.production


#install.packages("forecast")
library(forecast)

MO=c(2,1,1)
fit = Arima(Wt,order=MO,include.constant=TRUE,method = "ML")
fit

Est=coef(fit)
SE=sqrt(diag(fit$var.coef))
```

```
Tratio=Est/SE
Tabcd = c(Est,SE,Tratio)
TabCoef=matrix(Tabcd,I(MO[1]+MO[3]+1),3,dimnames = list(NULL,c("Est", "SE", "t")))
TabCoef
# Variance estimate
SSE = t(fit$residuals)%*%fit$residuals
# SSE/(length(Wt)-I(MO[1]+MO[2]+MO[3]+1))
# Eliminate "+1" if "constant=FALSE"

# Correlation Matrix
diag(1/SE,nrow=I(MO[1]+MO[3]+1))%*%fit$var.coef%*%diag(1/SE,nrow=I(MO[1]+MO[3]+1))


# Figs 7-5
par(mfrow=c(2,1))
Acf(residuals(fit),ylim=c(-1,1))
Pacf(residuals(fit),ylim=c(-1,1))
par(mfrow=c(1,1))

# Results not shown
# normal plot, periodogram
# fhat_5 spectral density
qqnorm(residuals(fit),ylab="Wt residuals")
spec.pgram(residuals(fit),taper=0,detrend=FALSE,fast=FALSE)
spec.pgram(residuals(fit),taper=0,detrend=FALSE,fast=FALSE,
  kernel("daniell", 2))

# Fig 7-6
plot(Year,residuals(fit),type="b")


# Fig 7-7


fitpred=forecast(fit,7)
fitpred

Yearfuture=c(81,82,83,84,85,86,87)
Wtfuture=c(818.4,833.5,777.9,891.8,879.0,886.1,912.7)
plot(Year,Wt,type="l",xlim=c(20,87),ylim=c(300,925))
lines(Yearfuture,Wtfuture,type="l")
lines(Yearfuture,fitpred$mean,type="l",lty=5)
lines(Year,fit$fit,type="l",lty=5)
legend("topleft",c("Actual","ARIMA"),lty=c(1,5))
```

6.6 Exercises

EXERCISE 6.6.1 Dodson (1995) looks at "daily viscosity readings from an oil house that supports a sheet metal rolling mill." The data are in Table 6.2.

regress on time and look at residuals for time series effect.

Table 6.2: *Oil house viscosity readings*

304	314	307	315	297	304	314	321	307	321
288	302	308	341	300	287	280	281	285	281
283	287	281	285	291	281	283	284	291	278
280	263	254	261	270	254	260	259	260	263
268	270	253	253	247	247	256	256	260	237
Read across then down									

Table 6.3: *Finished product carbon black DBP values*

pm			am			pm			am		
3	7	11	3	7	11	3	7	11	3	7	11
95	100	104	105	111	99	100	108	101	105	113	114
109	110	100	104	97	127	112	106	103	100	98	102
103	103	98	103	102	106	93	98	108	111	107	104
78	102	112	95	96	90	91	97	90	90	90	100
97	97	93	98	101	98	94	98	100	94	97	103
97	98	99	101	99	102	105	106	107	105	103	116

autoregress using least squares and look at resids for white noise
autoregress using ts procedure and look at resids
time series plot shows downward trend. compute first differences and replot. fit a model.

EXERCISE 6.6.2. Schneider and Pruett (1994, qe94-348.dat) look at carbon black DBP values over 12 days with 6 readings per day at 3, 7, and 11, both AM and PM (days start at 3PM). The data are in Table 6.3. Do a time series analysis. Look for day effects and for time within day effects.

EXERCISE 6.6.3. Consider again the data of Exercise 4.9.17 and Table 4.16.
The data are repeated in Table 6.4. Look for autocorrelations within Heads.

Table 6.4: *Outside diameter of injection molded bottles. (qe94-347.dat)*

	Head			
Sample	1	2	3	4
1	2.01	2.08	2.08	2.04
2	1.97	2.03	2.09	2.10
3	2.03	2.09	2.08	2.07
4	1.96	2.06	2.07	2.11
5	1.94	2.02	2.06	2.11
6	2.01	2.03	2.07	2.11
7	2.00	2.04	2.09	2.06
8	2.01	2.08	2.09	2.09
9	2.01	2.00	2.02	2.07
10	2.01	1.96	2.08	2.11
11	1.99	1.99	2.09	2.11
12	1.98	2.02	2.03	2.08
13	1.99	1.98	2.05	2.04
14	2.01	2.05	2.07	2.08
15	2.00	2.05	2.06	2.06
16	2.00	2.00	2.08	2.14
17	2.01	2.00	2.05	2.15
18	2.03	2.09	2.11	2.12
19	1.99	2.10	2.09	2.09
20	2.01	2.01	2.01	2.11

EXERCISE 6.6.4. Reanalyze all of the individuals chart data from the Chapter 4 exercises as time series.

Chapter 7

Reliability

7.1 Introduction

The topic of reliability covers a wide spectrum of material. Here we introduce the analysis of time-to-event data. Time-to-event data involves measuring how long it takes for something to happen: a ball bearing failure, an automobile crash, death due to cancer, etc. Analysis includes modeling and estimation of lifetime distributions and probabilistic methods for predicting and estimating mean lifetimes, survival probabilities, and survival times for both individual components and systems. In particular, we introduce some peculiarities that arise with time-to-event data

The first peculiarity of time to event data is some some new notation and concepts. Typically, we are concerned with the cumulative distribution function F and the probability density function f of random variables. With time to event data one is often interested in corresponding concepts, the reliability function R (known in biological applications as the survival function S) and the hazard function h. If you are testing how long a refrigerator lasts under stress, you may have to quit the study before all the refrigerators die. This creates censored data. Censored data seems much more prevalent when dealing with biological applications than industrial, but it raises its ugly head in both. These new ideas are introduced in Section 2.

The second peculiarity is that time to event data are often skewed rather than symmetric. As a result, statistical analysis is often based on distributions other than the normal. Several of these distributions are introduced in Section 3.

In general, reliability deals with the study of the proper functioning of equipment and systems. Our main area of focus is component and system reliability, i.e., in finding the reliability of a complex system from the knowledge of the reliabilities of its components. For a more extensive introduction to the subject, see Hoyland and Rausand (1994) or Crowder et al. (1991). For an alternative approach based on Bayesian statistics see Hamada et al. (2008).

In addition to the study of lifetime or failure-free operating time for a system or piece of equipment, reliability can include broader aspects of a systems performance over time. This allows varying levels of performance, possibility of repeated failures/repairs etc. It uses stochastic processes. e.g. Markov processes, to model system performance, cf. Huzurbazar (2004).

7.2 Reliability/Survival, Hazards and Censoring

7.2.1 Reliability and Hazards

Failure models are probability distributions used for modeling the time to failure of a component (or system) in reliability. Let T be the random time to failure of a unit, that is, the time from when the unit is put into operation until it first fails. For example, suppose T represents the time to breakage of a timing belt on a car. We could measure time here as calendar time but in this case a more relevant measure of "time" is the miles driven by the car. Here T represents the number of miles driven by car before the belt breaks and $t = 0$ is the mileage on the car when the belt was put into operation. For the remainder of this chapter, we assume that T is continuously distributed on $[0, \infty)$

with probability density function $f(t)$ and cumulative distribution function (cdf)

$$F(t) = \Pr(T \le t) = \int_0^t f(x)dx \qquad \text{for } t > 0. \tag{7.2.1}$$

$F(t)$ denotes the probability that the unit fails in the time interval $(0,t]$. By the Fundamental Theorem of Calculus, the probability density function is

$$f(t) = \frac{d}{dt}F(t). \tag{7.2.2}$$

The *reliability function* of a unit is defined as

$$R(t) \equiv 1 - F(t) = P(T > t). \tag{7.2.3}$$

This is the probability that the unit does not fail in the time interval $(0,t]$ or, equivalently, the probability that the unit is still functioning (survives) beyond the time t. In biological applications $R(t)$ is often called the *survival function* and denoted $S(t)$.

The probability that a unit fails in a short time, say, $(t,t+\delta t]$ is

$$\Pr(t < T \le t + \delta t) \doteq f(t)\delta t \tag{7.2.4}$$

Often, we are interested in the *hazard function* or *failure rate function*, denoted $h(t)$, which is defined from the probability that the unit fails in a short time interval $(t,t+\delta t]$ given that it has not failed by time t, i.e.,

$$\Pr(t < T \le t + \delta t | T > t) = \frac{F(t+\delta t) - F(t)}{1 - F(t)} \doteq \frac{f(t)\delta t}{[R(t)]}. \tag{7.2.5}$$

Mathematically, the hazard is the rate at which this probability changes at time t, that is, the limit as δt gets small of the conditional probability of the interval, divided by the length δt of the interval,

$$h(t) \equiv \lim_{\delta t \to 0} \frac{\Pr(t < T \le t + \delta t | T > t)}{\delta t} = \lim_{\delta t \to 0} \frac{F(t+\delta t) - F(t)}{[1 - F(t)]\delta t} = \frac{f(t)}{R(t)}.$$

The density function $f(t)$ and the hazard function $h(t)$ are mathematically interchangeable in that knowing one determines the other but they answer different questions. Suppose at time $t = 0$ we have a new unit and want the probability that this unit will fail in the interval $(t,t+\delta t]$. This is approximately equal to the probability density $f(t)$ times the length of the interval δt as given by (7.2.4). Now suppose that the unit has survived up to time t and we want to know the probability that this unit will fail in the next interval $(t,t+\delta t]$. This probability is now the conditional probability in (7.2.5) and is approximately equal to the hazard function $h(t)$ times the length of the interval δt.

If the hazard $h(t)$ is an increasing function of t, the chance of a failure becomes greater as time goes on. Such hazards are useful when the unit suffers from fatigue or cumulative damage. If the hazard function is a decreasing function of t, the chance of a failure decreases with time. This corresponds to the "burn-in" period for some electronic components where substandard units burn out early. If one restricts attention to an item's mature period, the chance of failure is often constant over time, i.e., the hazard is flat. Household appliances often display this when they are neither very young or old.

A bathtub shaped hazard function is often appropriate for the entire life of an item. Newly manufactured items (refrigerators, automobiles, even infants) are subject to high initial hazards that decrease as they mature. (If something is fundamentally wrong, it tends to show up early.) The items then settle into a long period of maturity where the hazard is essentially constant. Finally, as the items age or wear out, their hazards tend to increase again.

Figure 7.1 gives a bathtub hazard function showing the "burn-in", useful life, and wear-out period of a component. This hazard is a combination of the three shapes discussed previously. The

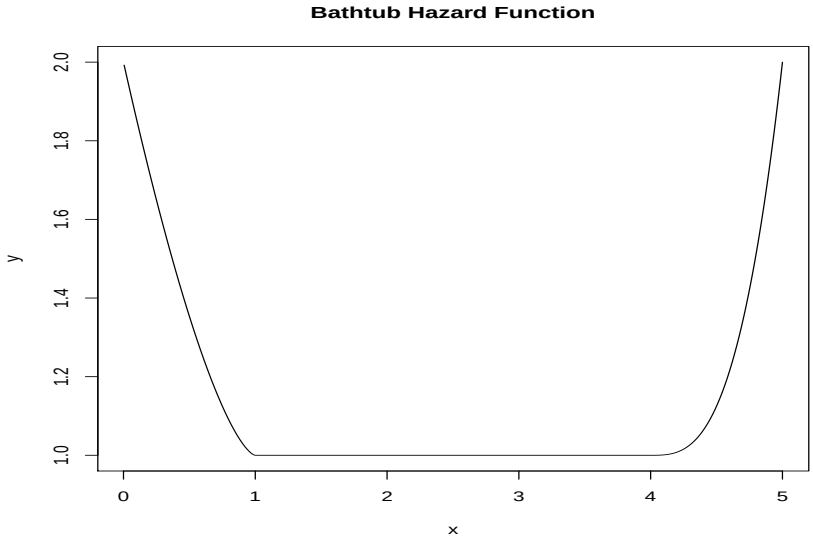

Figure 7.1: *Bathtub shaped hazard function.*

"burn-in" or "infant mortality" period shows a high hazard initially which represents undiscovered defects in substandard units. The "useful life" period is also called the "chance failure" period. This is the stable part of the unit's life. As the unit ages, it enters it's wear-out period. In practice, this hazard shape is difficult to use. Many units are tested via a quality control process before being released to users so that the "infant mortality" period is removed. For mechanical units, the useful life period often shows a slightly increasing tendency.

Often when analyzing failures we are restricted to limited age ranges of an item where the hazard shapes are constant, increasing, or decreasing. The Exponential distribution of the next section is often useful because it has a flat hazard. Many of the models in the next section display either flat or increasing hazards. Many include the Exponential as a special case. To have a flat hazard you must have an Exponential distribution.

The *mean time to failure (MTTF)* of a unit is defined as

$$MTTF = \mathrm{E}(T) = \int_0^\infty t f(t) dt$$

where $\mathrm{E}(T)$ is the expected value of the random variable T. The MTTF can also be expressed as

$$MTTF = \int_0^\infty R(t) dt.$$

In more complicated systems we might incorporate the time to repair an item, say, U. The *mean time between failures (MTBF)* is the MTTF plus the *mean time to repair (MTTR)*, i.e..

$$MTBF = MTBF + MTTR = \mathrm{E}(T) + \mathrm{E}(U).$$

7.2.2 Censoring

Consider the joint distribution of T, the time to event, and a random variable C the determines the time after which you will no long be able to observe T. If you started an experiment in the morning and have to quit and go home at 5PM, that determines C. In studies on people, C is when a person

becomes no longer available to the study. Perhaps they died. Perhaps they moved to another town. Perhaps they tired of the hassle of being studied. Censoring is particularly convenient when it occurs independently of the time T being studied. But if you think about the event being death, the closer death, gets the more likely one is to quit a study. When subject to censoring, the actual data observed are

$$(X,\delta) \equiv \begin{cases} (T,1) & \text{if } T \le C \\ (C,0) & \text{if } C > T. \end{cases}$$

δ is a random indicator for *not* being censored. (That does not keep people from sometimes calling it the censoring indicator.) Let f_T be the density for T and f_C the density for C. The likelihood for observing $X = u$ and δ is defined as

$$f_T(u)^{\delta} S_T(u)^{1-\delta}.$$

7.3 Some Common Distributions

Commonly used distributions for modeling the skewed data associated with time to event data are the Exponential, Lognormal, and Weibull. Another useful distribution is the Gamma.We now introduce these.

7.3.1 Exponential Distributions

The exponential distribution is the most widely used in reliability mainly because of its tractability and nice properties. We have already mentioned that it has a flat hazard function. It also has a memoryless property that

$$\Pr(T > t+u | T > t) = \Pr(T > u)$$

and the minimum of several iid Exponentials is again an Exponential (but with a different parameter).

The are two different parameterizations for the Exponential distribution that are in common use. The density, cdf, reliability, and hazard functions are given by

$$\begin{aligned} f(t) &= \theta \exp(-\theta t) \quad \text{for } \theta > 0,\ t > 0, \\ F(t) &= 1 - \exp(-\theta t), \\ R(t) &= \exp(-\theta t), \\ h(t) &= \frac{\theta \exp(-\theta t)}{\exp(-\theta t)} = \theta, \end{aligned} \tag{7.3.1}$$

or by

$$\begin{aligned} f(t) &= (1/\lambda) \exp(-t/\lambda) \quad \text{for } \lambda > 0,\ t > 0, \\ F(t) &= 1 - \exp(-t/\lambda), \\ R(t) &= \exp(-t/\lambda), \\ h(t) &= \frac{(1/\lambda) \exp(-t/\lambda)}{\exp(-t/\lambda)} = 1/\lambda. \end{aligned} \tag{7.3.2}$$

We will use the first parameterization and indicate T having an exponential distribution by $T \sim \text{Exp}(\theta)$. Note that if $X \sim \text{Exp}(1)$ then $T \equiv X/\theta \sim \text{Exp}(\theta)$.

7.3.2 Weibull Distributions

Weibull distributions are a straightforward generalization of the exponential. If $X \sim \text{Exp}(\theta)$ then $T = X^{1/\alpha}$ has a Weibull(α,θ) distribution when the Weibull is parameterized as

$$f(t) = \alpha\theta t^{\alpha-1} \exp(-\theta t^{\alpha}) \quad \text{for } \theta > 0,\ \alpha > 0,\ t > 0$$

$$\begin{aligned} F(t) &= 1-\exp(-\theta t^{\alpha}) \\ R(t) &= \exp(-\theta t^{\alpha}) \\ h(t) &= \alpha\theta t^{\alpha-1}. \end{aligned}$$

An alternate parameterization is

$$\begin{aligned} f(t) &= \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}\exp[(-t/\lambda)^{\alpha}] \quad \text{for } \lambda>0,\ \alpha>0,\ t>0 \\ F(t) &= 1-\exp[(-t/\lambda)^{\alpha}] \\ R(t) &= \exp[(-t/\lambda)^{\alpha}] \\ h(t) &= \left(\frac{\alpha}{\lambda^{\alpha}}\right)t^{\alpha-1}. \end{aligned}$$

For $\alpha = 1$ this is the exponential. For $\alpha > 1$ the hazard function is increasing as a power of t. For $\alpha < 1$ the hazard is decreasing.

7.3.3 Gamma Distributions

Gamma distributions are a less straightforward generalization of the exponential.

The are two different parameterizations for the Gamma distribution that are in common use.

$$f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}t^{\alpha-1}e^{-\beta t} \quad \text{for } \alpha>0,\ \beta>0,\ t>0$$

or

$$f(t) = \frac{1}{\Gamma(\alpha)\lambda^{\alpha}}t^{\alpha-1}e^{-t/\lambda} \quad \text{for } \alpha>0,\ \lambda>0,\ t>0.$$

There are no particularly nice functional forms for the cdf, reliability function, or hazard. α is known as the shape parameter, β is a scale parameter.

Using the Gamma(α,β) parameterization, a Gamma$(1,\theta)$ is an Exp(θ).

7.3.4 Lognormal Distributions

If $X \sim N(\mu,\sigma^2)$ then $Y = \log(X)$ has a lognormal distribution with median e^{μ} and scale parameter σ. Turns out the the mean is $\exp[\mu + (\sigma^2/2)]$.

7.3.5 Pareto

This is primarily of interest because it has a decreasing hazard function. It can be constructed as a mixture of exponential distributions using gamma distribution weights. (To a Bayesian, it is the marginal distribution of the data for a conditional distribution that is exponential and a prior that is gamma.)

$$\begin{aligned} f(t) &= \frac{\beta\alpha^{\beta}}{(t+\alpha)^{\beta+1}} \quad \text{for } \alpha>0,\ \lambda>0,\ t>0, \\ F(t) &= 1-\left(\frac{\alpha}{t+\alpha}\right)^{\beta}, \\ R(t) &= \left(\frac{\alpha}{t+\alpha}\right)^{\beta}, \\ h(t) &= \beta/(t+\alpha). \end{aligned}$$

Table 7.1: *Plackett-Burman L_{12} Screening design for thermostats*

Trial	A	B	C	D	E	F	G	H	I	J	K
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	1	1	1	1
3	0	0	1	1	1	0	0	0	1	1	1
4	0	1	0	1	1	0	1	1	0	0	1
5	0	1	1	0	1	1	0	1	0	1	0
6	0	1	1	1	0	1	1	0	1	0	0
7	1	0	1	1	0	0	1	1	0	1	0
8	1	0	1	0	1	1	1	0	0	0	1
9	1	0	0	1	1	1	0	1	1	0	0
10	1	1	1	0	0	0	0	1	1	0	1
11	1	1	0	1	0	1	0	0	0	1	1
12	1	1	0	0	1	0	1	0	1	1	0

Table 7.2: *Number of temp cycle changes before thermostat failure.*

Trial	Data
1,	, 957, 2846, 7342, 7342, 7342, 7342, 7342, 7342, 7342, 7342,
2,	, 206, 294, 296, 305, 313, 343, 364, 420, 422, 543,
3,	, 63, 113, 129, 138, 149, 153, 217, 272, 311, 402,
4,	, 76, 104, 113, 234, 270, 364, 398, 481, 517, 611,
5,	, 92, 126, 245, 250, 390, 390, 479, 487, 533, 573,
6,	, 490, 971, 1615, 6768, 7342, 7342, 7342, 7342, 7342, 7342,
7,	, 232, 326, 326, 351, 372, 446, 459, 590, 597, 732,
8,	, 56, 71, 92, 104, 126, 156, 161, 167, 216, 263,
9,	, 142, 142, 238, 247, 310, 318, 420, 482, 663, 672,
10,	, 259, 266, 306, 337, 347, 368, 372, 426, 451, 510,
11,	, 381, 420, 7342, 7342, 7342, 7342, 7342, 7342, 7342, 7342,
12,	, 56, 62, 92, 104, 113, 121, 164, 232, 258, 731

7.4 An Example

Bullington, Lovin, Miller, and Woodall (1993) examine the number of temperature cycle changes a thermostat endures before failure. They examined 11 factors each at 2 levels that were thought to possibly affect the performance.

A: Diaphram Plating Rinse (clean primary – pure secondary, contaminated primary – no secondary)
B: Current Density (5 min at 60 amps, 10 min at 15 amps)
C: Sulfuric Acid Cleaning (3 sec, 30 sec)
D: Diaphragm Electroclean (2 min, 12 min)
E: Beryllium Copper Grain Size (0.008in, 0.018in)
F: Stress Orientation (perpendicular, in line – relative to seam weld)
G: Humidity (wet, dry)
H: Heat Treatment (45 min, 4hrs – at 600° F)
I: Brazing Machine (no cooling water or flux – excess water, 3 times normal flux)
J: Power Element Electroclean (like D)
K: Power Element Plating Rinse (like A)

This determines $2^1 1$ treatment combinations of which only 12 were observed as indicated in Table 7.1 wherein 0s indicate the first factor level as given above.

Table 7.1 is a Placket-Burman design as discussed in *TiD* and a Taguchi L_{12} design as discussed in Chapter 11 and *TiD*. Table 7.2 contains the 10 observations on each observed treatment combinations.

Rather than doing a full analysis of this complicated design, use one treatment to illustrate a one

sample problem and two treatments to illustrate two sample problems. Introduce censoring from some trials.

Some Random Thoughts

give an example early on to motivate these things. why does anyone really care about R(t) or h(t)? The first couple of things are bullets to be worked into an example

- Nuclear reactor safety study. Need to look up.
- Germany WWII; V1 missile. First 10 were fiascos. Careful attention was paid to details but first 10 exploded on the launch pad or landed "too soon" in the English channel. Later, people thought about probability for series components.
  P(system works)= $\prod$ P(individual components work)¡¡ P(any one component works)

  Imagine we have 10 components in series. Each has prob 0.90 of working. Prob the system works is $.9^{10} = .3486$ which is far less than the prob of any individual component working.
- Point out that this will only get worse when we have components in series and parallel as is the case with a full-fledged missile or even your washing machine...

start with what is reliability in general

Things to add possibly

- repairable systems. k out of p systems, system functions when $\geq$ k components work properly.

Chapter 8

Random Sampling

Random sampling is the tool used to ensure that samples are collected without bias. It is all too easy for people to choose samples that will bias the results of the study. If one is studying how much milk school children drink in a school district, a sample chosen by the whim of teachers would probably be biased. The students chosen would probably not be representative of the students as a whole. In fact, they may well be chosen (perhaps subconsciously) to reinforce the attitudes of the teacher.

If a shipment of automobile components were being evaluated to check the rate of defectives, the containers at the front of the shipment may not be representative of the entire shipment. This could happen if the supplier is disreputable and suspects that you will only sample those crates that are easy to sample. More realistically (we hope) the state of the production process can change over time and so when crates are placed for shipment, they are placed in time order, so where you sample is related to the time of production. Random sampling is a device for selecting a sample that has no systematic bias.

8.1 Acceptance Sampling

Suppose you manufacture audio equipment and you have just received a shipment of electrical components. How do you decide whether these components are of sufficiently high quality for your manufacturing process?

By far the best way to do this is to have a history with the supplier which tells you that they are a reputable firm. Then, have the supplier show you the control charts they used in production. Based on the control charts, if the manufacturing process was under control and meeting specifications, you can purchase the item without reservations.

What if you don't trust the supplier to show you accurate control charts? Why are you dealing with a supplier like that? What if this is a new supplier that does not have control charts? Why are you dealing with a supplier like that? We could go on and on like this, but the fact is that sometimes people end up making purchases under less than optimal conditions. In such situations, how can you decide whether to accept a shipment of components or not. A reasonable method is to take a random sample of the shipment and inspect the sample to learn about the shipment as a whole.

Deming would probably argue that while sampling may be reasonable to determine what to do with the entire shipment, once the decision is made to use the components, sampling should either cease or every single component should be inspected to determine if it meets specifications. Juran and Gryna (1993) give reasons for performing acceptance sampling but they suggest that acceptance sampling should be used in an overall quality control process designed to eliminate the need for acceptance sampling. Some situations where sampling is important are destructive testing (in which to test a product you have to destroy it), expensive testing (in which the cost of examining every unit is excessive), situations where production is continuous, so that no natural definition of a manufacturing unit exits (examples of this are textiles, thread, paper, photographic film, etc.) and where production runs are too large and too fast, so that there are too many units to inspect (examples include the production of nuts, bolts, screws, etc.).

8.2 Simple Random Sampling

Suppose a shipment of automobile components contains $N = 300$ units. Suppose we want to take a sample of $n = 50$ of these units to inspect for defectives. To take a simple random sample, we have to make a list of all 300 units, assigning each a number between 1 and 300. It doesn't matter how we make the list as long as each unit appears once and only once on the list. It could be a list of serial numbers or, if the components were delivered on a palate in crates, one to a crate, we can just imagine assigning a number to each crate on the palate. Using a method for generating random numbers, you randomly select 50 numbers between 1 and 300 without replacement. Having obtained the 50 numbers, go back to the list and identify the automobile components that correspond to the numbers on the list. These are the automobile components that need to be inspected for defects.

The Minitab commands below will make such a selection and list the 50 numbers in column C2.

```
set c1
1:300
end.
sample 50 c1 c2
```

Or you can enter the same information by navigating the menus:

```
Calc > Make Patterned Data > Simple Set of Numbers
Calc > Random Data > Sample from Columns
```

To fix ideas we present an artificially small example.

EXAMPLE 8.2.1 A collection of $N = 40$ video tapes, I mean DVDs, I mean downloads, no I really do mean DVDs, of the popular English costume comedy *Blue Asp* were obtained. The running times of these DVDs are designed to be exactly 34 minutes. Each DVD is clearly identified and assigned a number from 1 to 40. A simple random sample of size $n = 5$ from the numbers between 1 and 40 gave 9, 23, 13, 38, 6. The 5 DVDs corresponding to these numbers were pulled out of the collection and the running times of the DVDs were measured. The run times were 34.610, 34.743, 34.480, 34.239, 33.939. The estimate of the population mean running time is the sample mean running time

$$\hat{\mu} \equiv \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{34.610 + 34.743 + 34.480 + 34.239 + 33.939}{5} = 34.4022$$

The estimate of the population variance is the sample variance

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{5-1}\left[(34.610 - 34.4022)^2 + \cdots + (33.939 - 34.402)^2\right] \\ &= 0.1016417 \end{aligned}$$

The **usual** standard error of $\bar{x}$ from an introductory statistics course is

$$\mathrm{SE}(\bar{x}) = \sqrt{\frac{s^2}{n}}$$

with a 95% confidence interval for the population mean running time of

$$\bar{x} \pm 2\,\mathrm{SE}(\bar{x}).$$

Alas, *a slight modification of the standard error is necessary in our current situation*. The usual results given above assume that either the population being sampled is infinitely large or that the sampling occurs with replacement. Sampling with replacement means that after an object has been sampled from the population, it is placed back into the population so that it has the possibility of

being sampled again. That is **not** how simple random sampling is conducted in practice. In practice, sampling is without replacement. Once an object has appeared in a sample, it cannot appear again. This actually causes the sampling to be more accurate. If you take a simple random sample with replacement of size 5 from a population of size 5, you may or may not see every element in the population. If you take a simple random sample without replacement of size 5 from a population of size 5, you are guaranteed to have seen every element in the population and the sample mean is guaranteed to equal the population mean. Similarly, there will be less variability in $\bar{x}$ for a sample of size 5 from a population of size 40 when sampled without replacement, than there will be if the population is sampled with replacement. The difference can be accounted for by a finite population adjustment,

$$1-\frac{n}{N}$$

where n is the sample size and N is the population size. The sample mean and variance without replacement remain

$$\hat{\mu} \equiv \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = 34.4022$$

and

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= 0.1016417 \end{aligned}$$

but now, using the finite population adjustment,

$$\mathrm{SE}(\bar{x}) = \sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}} = \sqrt{\left(1-\frac{5}{40}\right)\frac{0.1016417}{5}} = 0.13337.$$

With this new standard error, the 95% confidence interval is

$$\bar{x} \pm 2\,\mathrm{SE}(\bar{x})$$

or

$$34.4022 \pm 2(0.13337)$$

for in interval of (34.135,34.669). The interval does not contain the target value 34, so the DVDs seem to be running longer than they are supposed to run.

It turns out that for sampling without replacement

$$\mathrm{E}(\bar{x}) = \mu, \quad \mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right), \quad \mathrm{E}(s^2) = \sigma^2\frac{N}{N-1}.$$

From these it is not difficult to show that

$$\mathrm{E}\left[\left(1-\frac{n}{N}\right)\frac{s^2}{n}\right] = \mathrm{Var}(\bar{x}).$$

Note that, whenever n approaches N, $\mathrm{Var}(\bar{x})$ approaches 0 and that, as N gets large, these approach the usual sampling with replacement results.

8.3 Stratified Random Sampling

Suppose you have bought automobile components from a company, but the parts delivered come from three different manufacturing plants. For production purposes, it is probably best to just treat

these as three different shipments and evaluate the quality of the three shipments separately. However for some purposes, perhaps cost accounting, it may be necessary to get an estimate of the overall total number of defectives coming from this company. The appropriate technique is to take a simple random sample of components from each of the plants and to combine the results appropriately. In this context, each of the plants is referred to as a strata and taking random samples from each strata is referred to as stratified random sampling.

EXAMPLE 8.3.1 In addition to the *Blue Asp* DVDs obtained in Example 8.2.1, two other shipments of DVDs were obtained at a later time, each of which included 30 DVDs. We want to estimate the population mean of all 100 DVDs. We begin by taking simple random samples without replacement from each of the three shipments (strata). For illustrative purposes, the sample from the first shipment is as in Example 8.3.1 and the samples from the other two shipments are of sizes 4 and 6, respectively. The samples are

Stratum	N_i	Sample
1	40	34.610, 34.743, 34.480, 34.239, 33.939
2	30	34.506, 34.734, 34.350, 34.509
3	30	34.016, 34.347, 34.084, 34.932, 34.605, 34.437

Summary statistics are given below. Subscripts are used to indicate the stratum associated with a stratum size, a sample size, a sample mean, or a sample variance. For example, $N_1 = 40$, $n_1 = 5$, $\bar{x}_1 = 34.4022$, and $s_1^2 = .1016417$ are all from stratum 1.

Stratum	N_i	n_i	$\bar{x}_i$	s_i^2
1	40	5	34.4022	0.1016417
2	30	4	34.52475	0.02497425
3	30	6	34.4035	0.1152931

The estimate of the mean of the entire population is a weighted average of the stratum sample means. The weights are relative to the size of the strata N_i. Let

$$N = N_1 + N_2 + N_3 = 40 + 30 + 30 = 100$$

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{3} \frac{N_i}{N} \bar{x}_i \\ &= \frac{40}{100} 34.4022 + \frac{30}{100} 34.52475 + \frac{30}{100} 34.4035 \\ &= 34.439355 \end{aligned}$$

The variance of the estimate is

$$\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{3} \left(\frac{N_i}{N} \right)^2 \mathrm{Var}(\bar{x}_i)$$

so, using the results on simple random sampling for each strata (which includes incorporating the finite population corrections for sampling within each strata), the standard error is,

$$\begin{aligned} &\mathrm{SE}(\hat{\mu}) \\ &= \sqrt{\sum_{i=1}^{3} \left(\frac{N_i}{N} \right)^2 \left(1 - \frac{n_i}{N_i} \right) \frac{s_i^2}{n_i}} \\ &= \left[\left(\frac{40}{100} \right)^2 \left(1 - \frac{5}{40} \right) \frac{0.1016417}{5} + \left(\frac{30}{100} \right)^2 \left(1 - \frac{4}{30} \right) \frac{0.02497425}{4} \right. \\ &\quad \left. + \left(\frac{30}{100} \right)^2 \left(1 - \frac{6}{30} \right) \frac{0.1152931}{6} \right]^{1/2} \\ &= 0.068676653 \end{aligned}$$

A 95% confidence interval for the mean has endpoints $\hat{\mu} \pm 2\text{SE}(\hat{\mu})$, which is

$$34.439355 \pm 2(0.068676653)$$

giving the interval (34.302, 34.577).

8.3.1 Allocation

One question about stratified samples is how do you decide how many samples to take in each stratum. One relative safe way, is to use proportional allocation of the samples. In proportional allocation, for an over all sample size of n, you choose n_i observations from stratum i where n_i is chosen to satisfy

$$\frac{n_i}{n} = \frac{N_i}{N}$$

Note that the total number of units in the ith stratum, N_i, and the total number of units in the population, N, are both assumed to be known.

If you have some idea about how much it costs to collect an observation in each stratum and if you also have some idea of the variance in each stratum, one can find an optimal allocation. In optimal allocation, you will want to take more observations in strata with more variability and fewer observations in strata for which the cost per observation is more. These issues are discussed in more detail in any good book on sampling, e.g., Lohr (1999).

8.4 Cluster Sampling

Suppose again that the automobile components have been crated and put on a palette, but that now these are small components and that there are 20 components in each crate. To simplify life, lets assume that they all come from the same plant. The obvious way to sample is to make a list of crates and take a simple random sample of crates. Once the crates have been selected, either we can inspect all 20 of the components in each crate, or for each crate we can take a further random sample of the 20 components and inspect only those components chosen in the second stage of sampling. The first of these procedures is referred to as one-stage cluster sampling, the second procedure is referred to as two-stage cluster sampling.

EXAMPLE 8.4.1 *One-stage cluster sampling*.
Suppose *Blue Asp* DVDs are boxed in cartons of 20 to 25 and a wholesaler gets a shipment of 100 cartons. Suppose the wholesaler wants to estimate the population mean running length of a DVD based on taking a sample of 3 cartons. This is a very small sample but it will illustrate the ideas. This involves sampling 3 cartons and measuring the running time for each DVD in the carton. Implicitly we are assuming that DVDs in a carton are probably more alike that DVDs in different cartons.

Carton	N_i	$\bar{x}_i$
1	20	34.413
2	23	34.535
3	25	34.414

Note that both the carton sizes and the carton means are random variables but only because we have a random selection of cartons. □

In general let c denote the number of clusters (cartons) in the population. Let the pair (N_k, T_k) denote the number of items in the kth cluster and the total of all the measurements made on the items in the kth cluster. (The total might be the total number of defectives in the cluster.) Note $T_k = N_k \bar{x}_k$ where $\bar{x}_k$ is computed from the x_{kj}s involving the jth observation from the kth cluster.

The population mean of the x_{kj} population is the sum of all the cluster totals T_k divided by the sum of all the cluster sizes N_k, i.e.,

$$\mu = \frac{\sum_{k=1}^{c} T_k}{\sum_{k=1}^{c} N_k} = \frac{\sum_{k=1}^{c} N_k \bar{x}_k}{\sum_{k=1}^{c} N_k} = \frac{\sum_{k=1}^{c} \sum_{j=1}^{N_k} x_{kj}}{\sum_{k=1}^{c} N_k}. \tag{8.4.1}$$

The odd thing about cluster sampling is that we are not directly sampling from our population of interest, the x_{kj}s. And, unlike our previous problems, we do not even know the size of the x_{kj} population. We know the number of clusters c but we do not know the size of every cluster. This means that, unlike our previous scenarios, we need to estimate both the numerator and denominator of equation (8.4.1).

Take a random sample of n clusters giving us random pairs (N_i, T_i), $i = 1, \ldots, n$. The mean cluster size is

$$\bar{N} = \frac{\sum_{i=1}^{n} N_i}{n},$$

so the estimated x_{kj} population size is $c\bar{N}$. The estimated mean cluster total is $\bar{T}$ so the estimated x_{kj} total is $c\bar{T}$. The estimated x_{kj} population mean is

$$\hat{\mu} = \frac{c\bar{T}}{c\bar{N}} = \frac{\bar{T}}{\bar{N}} = \frac{\sum_{i=1}^{n} T_i/n}{\sum_{i=1}^{n} N_i/n} = \frac{\sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} N_i} = \frac{\sum_{i=1}^{n} N_i \bar{x}_i}{\sum_{i=1}^{n} N_i}. \tag{8.4.2}$$

The usual standard error is approximated by treating the N_is as fixed so that all the randomness is associated with $\bar{T}$,

$$\mathrm{SE}(\hat{\mu}) = \mathrm{SE}\left(\frac{\bar{T}}{\bar{N}}\right) \doteq \frac{1}{\bar{N}} \mathrm{SE}(\bar{T}) = \frac{1}{\bar{N}} \sqrt{\left(1 - \frac{n}{c}\right) \frac{s_T^2}{n}}$$

where, again treating N_i as fixed so that $\mathrm{E}(T_i) = N_i \mu$,

$$s_T^2 = \frac{1}{n-1} \sum_{i=1}^{n} (T_i - N_i \hat{\mu})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (N_i \bar{x}_i - N_i \hat{\mu})^2 = \frac{1}{n-1} \sum_{i=1}^{n} N_i^2 (\bar{x}_i - \hat{\mu})^2.$$

EXAMPLE 8.4.1 CONTINUED. *One-stage cluster sampling.*
Applying the formulae to our DVD example gives.

$$\hat{\mu} = \frac{20(34.413) + 23(34.535) + 25(34.414)}{20 + 23 + 25} = 34.45463.$$

$$s_T^2 = \frac{[20^2 (34.413 - \hat{\mu})^2] + [23^2 (34.535 - \hat{\mu})^2] + [25^2 (34.414 - \hat{\mu})^2)]}{2} = 2.570979,$$

$$\mathrm{SE}(\hat{\mu}) = \frac{1}{22.6666666} \sqrt{\left(1 - \frac{3}{100}\right) \frac{s_T^2}{3}} = 0.04022415.$$

A 95% confidence interval is

$$34.45463 \pm 2(0.04022415).$$

or (34.37418, 34.53508).

The variance and standard error are ugly arithmetic so I wrote R code:

```
muhat=((20 *34.413) + (23 * 34.535) + (25  *34.414))/68
muhat
sT2=( (20^2 * (34.413-muhat)^2) + (23^2 * (34.535-muhat)^2) +
      (25^2  *(34.414-muhat)^2))/2
```

```
sT2
se=sqrt( (1-(3/100))*sT2/3  )/22.6666666
se
muhat+(2*se)
muhat-(2*se)
```

□

EXAMPLE 8.4.2 *Two-stage cluster sampling*.
Measuring the running length of each DVD in a carton of 20 seems excessive. It might be more reasonable to randomly select a smaller number of DVDs out of each carton and measure the running length on just those DVDs.

In our example we have $c = 100$ clusters from which a sample of $n = 3$ is taken. The size of each cluster in the sample is N_i and the size of the sample within each cluster is n_i. Summary statistics for our example of DVD lengths are:

Carton	N_i	n_i	$\bar{x}_i$	s_i^2
1	40	5	34.4022	0.1016417
2	30	4	34.52475	0.02497425
3	30	6	34.4035	0.1152931

In one-stage cluster sampling we looked at every DVD in the carton, so we did not need to consider s_i^2, the variance associated with sampling from (within) the carton □

In a technical sense, two-stage cluster sampling is very similar to stratified sampling in which each cluster is thought of as a stratum. The key difference is that in stratified sampling, we collect a sample from every stratum (cluster) and in two stage cluster sampling we only take a sample of the clusters (strata). As a result, the formulae for two-stage cluster sampling include both stratified sampling and one-stage cluster sampling as special cases. Stratified sampling is where we take a complete sample of all the clusters (strata) and one-stage cluster sampling is where the sample within the clusters contains all of the units within the cluster. Ideally, we would perform stratified sampling when there is a lot of variability between clusters and not much variability within them, and we would perform one-stage cluster sampling when there is little variability between clusters and a lot of variability within them, but, in practice, what we perform is usually determined by physical characteristics of the population.

Our population of interest is the x_{kh}s, $k = 1,\ldots,c$, $h = 1,\ldots,N_k$. Again, $T_k = \sum_{h=1}^{N_k} x_{kh}$ is the total from the kth cluster and just as in equation (8.4.1),

$$\mu = \frac{\sum_{k=1}^{c} T_k}{\sum_{k=1}^{c} N_k}.$$

Again, we need to estimate both the numerator and the denominator.

In two-stage cluster sampling we take a first-stage sample of n clusters from the population of c clusters The second-stage sample from the ith cluster involves determining N_i as well as taking a sample of size n_i, say x_{ij}, $j = 1,,\ldots,n_i$. To estimate the total in the ith cluster, we use $\hat{T}_i = N_i\bar{x}_i$. The estimated x_{kj} population mean is a minor modification of (8.4.2),

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \hat{T}_i}{\sum_{i=1}^{n} N_i} = \frac{\sum_{i=1}^{n} N_i\bar{x}_i}{\sum_{i=1}^{n} N_i}.$$

The estimated population mean can also be written as,

$$\hat{\mu} = \frac{1}{\bar{N}} \frac{\sum_{i=1}^{n} N_i\bar{x}_i}{n}. \tag{8.4.3}$$

The standard error now has to incorporate not only the variability due to looking at only n of

c clusters but also the variability of estimating the total T_i from the sampled clusters. The former involves a version of s_T^2 and the latter involves the within cluster sample variances s_i^2.

$$\mathrm{SE}(\hat{\mu}) = \frac{1}{\bar{N}}\sqrt{\left(1-\frac{n}{c}\right)\frac{s_T^2}{n} + \frac{1}{nc}\sum_{i=1}^{n}\left(1-\frac{n_i}{N_i}\right)N_i^2\frac{s_i^2}{n_i}}. \tag{8.4.4}$$

The formula

$$s_T^2 = \frac{1}{n-1}\sum_{i=1}^{n}(N_i\bar{x}_i - N_i\hat{\mu})^2$$

looks the same as for one-stage cluster sampling but the components $\bar{x}_i$ and $\hat{\mu}$ are computed differently in two-stage cluster sampling.

Sometimes the size is known for every cluster in the population. In that case, $\bar{N}$ should be replaced in (8.4.3) and (8.4.4) by $\tilde{N} \equiv (N_1 + \cdots + N_c)/c$, the population mean of the cluster sizes.

Note that if $n = c$, we get the formulae for stratified sampling and, if $n_i = N_i$ for all i, we get the formulae for one-stage cluster sampling. So these two-stage formula are really the only ones you need to know from this chapter!

EXAMPLE 8.4.2 CONTINUED. *Two-stage cluster sampling*.
Applying the formulae to the two-stage DVD data,

$$\hat{\mu} = \frac{40(34.4022) + 30(34.52475) + 30(34.4035)}{40+30+30} = 34.43935,$$

$$s_T^2 = \frac{[40^2(34.4022-\hat{\mu})^2] + [30^2(34.52475-\hat{\mu})^2] + [30^2(34.4035-\hat{\mu})^2]}{2} = 5.414398,$$

$$\begin{aligned}
&\sum_{i=1}^{n}\left(1-\frac{n_i}{N_i}\right)N_i^2\frac{s_i^2}{n_i} \\
&= \left(1-\frac{5}{40}\right)40^2\frac{0.1016417}{5} + \left(1-\frac{4}{30}\right)30^2\frac{0.02497425}{4} + \left(1-\frac{6}{30}\right)30^2\frac{0.1152931}{6} \\
&= 47.16483.
\end{aligned}$$

$$\mathrm{SE}(\hat{\mu}) = \left(\frac{1}{33.3333333}\right)\sqrt{\left(1-\frac{3}{100}\right)\frac{s_T^2}{3} + \frac{1}{3(100)}[47.16483]} = 0.04143772.$$

The 95% confidence interval is

$$34.43935 \pm 2(0.04143772)$$

or (34.35648, 34.52223). The estimate of the population size N is $c\bar{N} = 100(33.3333333) = 3333$.

Again I wrote R code:

```
muhat=((40 *34.4022) + (30 * 34.52475) + (30  *34.4035))/100
muhat
sT2=( (40^2 * (34.4022-muhat)^2) + (30^2 * (34.52475-muhat)^2) +
    (40^2  *(34.4035-muhat)^2))/2
sT2
sisum=( 1 - (5/40) )*40^2 *(0.1016417/5) +
  ( 1 - (4/30) )*30^2 *(0.02497425/4) + ( 1 - (6/30) )*30^2 *(0.1152931/6)
sisum
se=(1/33.3333333) * sqrt( (1-(3/100))*sT2/3 + (1/(3*100)) * sisum)
se
muhat+(2*se)
muhat-(2*se)
```

□

8.5 Final Comment.

There are many reasons for collecting accurate information about a body of material (population). The methods presented here are useful for these purposes. However, the worst reason for conducting a sample survey of a population is to determine the percentage of defective units in that population. The only acceptable proportion of defective units should be 0. Attention should be focused on product improvement, not the bean counting of how often things have gone wrong. Fixing the blame for defects does not fix the problem that created the defects. The blame properly belongs to a management system that allows things to go wrong. Quantifying the extent of the problem can only help when determining which problems are most important to fix first. Quantifying the problem also gives no help in identifying how to fix the problem. Other statistical procedures can help with that.

Chapter 9

Experimental Design

To find out the effect of changing a process, it is generally necessary to change the process. To find out if a chemical reaction works better at 200 degrees than it does at 190 degrees, you need to set the process up to work at 200 degrees and see what happens. Seeing what happens when the process strays up to 200 degrees may be a good reason to investigate behavior at 200 degrees, but there is no assurance that the process will work the same when set to 200 degrees as it did when it strayed to 200 degrees. There was some reason for the process straying to 200 degrees, and any number of factors other than the temperature could be responsible for the observed improvement. To be sure that the *cause* of the improvement is the change in temperature, an experiment must be run. In an experiment, material is usually selected with as little variability as possible (while still characteristic of typical operating conditions) and the treatments (temperatures) are randomly assigned to the material.

The quality of fiberglass shingles is tested using an Elmendorf tear test. TAMKO Roofing Products of Joplin, MO was interested in evaluating the testing procedure. For their purposes, they needed a material to test with very little intrinsic variability so that the observed variability would be due to the variability of the measurements rather than the variability of the material. It was eventually decided to test the tear strength of vinyl floor covering under specified conditions. Two 1 foot by 6 foot pieces were purchased and cut into five 1 foot squares. The squares were cut into 3-in by 2.5-in specimens. Four trained Elmendorf operators were each given a stack of 10 randomized specimens and instructed to test them in order. The experiment is well suited for identifying any differences in performance due to the four operators and estimating the variability associated with the measurements. In reality, this is a brief synopsis of a data collection regime presented in Phillips et al. (1998) for evaluating the variability of the tear testers.

Experimental data are the only data that can be reliably used to determine cause and effect. Observational data can be used to establish that groups are different but observational data cannot be used to establish *why* groups are different. For example, if one has data on the ages at which native americans, hispanics, and non-hispanic caucasians commit suicide, one may be able to establish that these groups differ in the average ages at which they commit suicide, cf. Koopmans (1987) and Christensen (1996, Chapter 5). However, knowing that there are differences in the groups does nothing to explain *why* there are differences in the groups. It does nothing to establish cause and effect. Statistical data analysis is incapable of establishing cause and effect. Other ideas beyond data analysis are necessary for establishing this.

Statistical experiments are characterized by using random numbers to assign experimental treatments to experimental material. It is this use of random numbers that logically allows one to infer cause and effect from statistical experiments. Using random numbers to assign treatments to material eliminates any systematic biases among the groups of material associated with a treatment. Since there can be no systematic differences between the groups of material to which the treatments were applied, any differences observed between the treatment groups must be caused by the treatments. Note that the data analysis only establishes that the groups are different, it is the experimental procedure that allows one to infer that the differences are caused by the treatments.

For example, suppose one were to study the effect of giving young school children a cup of

milk each day. In each of several classrooms, half the children are given a cup of milk and half the children are not. At the end of the school year, measures of student performance are taken. Left to their own devices, the classroom teachers may well give the milk to the half of their class that seem to "need" the milk most. In other words, they may give the milk to those students with the lowest socio-economic status. Most measures of student performance are highly correlated with socio-economic status, and a cup of milk each day is unlikely to overcome those differences. Thus, by giving the milk to the most needy students, the teachers are assuring that, as a group, students who are given a cup of milk each day will perform *worse* than students who do not receive milk. To accurately measure the effect of giving out a cup of milk each day, those receiving the milk must be chosen randomly, so that there are no systematic differences between the experimental groups other than the experimental treatments. Similar phenomena can easily occur in industrial experiments. In the tear testing example, the experimental material (pieces of vinyl floor covering) were prepared first and then randomly assigned to the operators. When assigning treatments in an experiment, most definitely, the squeaky wheel *does not* necessarily get the grease. *A true experiment requires randomly assigning treatments to experimental units. Anything else is an observational study.*

Great care must be used in setting up an experiment so that the treatments are actually what they were designed to be. Once the experimental material has been randomly assigned to treatment groups, a treatment becomes *everything* that systematically happens to that experimental group. Suppose you randomly assign rats to two groups and give one a drug and the other a placebo. If the rats in the "drug" group develop cancer at a much higher rate than the placebo group, does that imply that the drug caused the additional cancers? Not if you were storing the drug rats in an area contaminated by asbestos (and the placebo rats someplace else)! The "treatments" can logically be said to be causing the extra cancer, but the "drug" treatment is actually the combination of the drug and the asbestos and anything else that was systematically different for the "drug" group than for the "placebo" group. It is important not to confuse the treatment with the label for the treatment. Even more importantly, one needs to set up experiments so that the treatment is as close as possible to being what was originally meant by the treatment label.

Some conditions that one would like to study cannot be random assigned to experimental material, e.g., sex cannot be randomly assigned to people. So it is impossible to conduct a statistical experiment with sex as a treatment. Moreover, sexes are labels for such broad collections of traits that differ between two populations that it would be difficult to ever think of describing sex as the *immediate* cause of anything that would require an experiment to establish. Sex groups may be different, but what is it about being female that makes them behave or respond differently than males?

One final word needs to be said about randomization as the basis for concluding that treatments cause observed effects. Randomization really only applies over the long run. It is possible, though unlikely, that systematic differences will occur in the treatment groups as the result of randomization. It is possible, unlikely but possible, that all of the "good" experimental material will randomly get assigned to one treatment group. If you see that that has happened, then you should do something about it. Unfortunately, if you don't know what constitutes "good" material, you cannot be cognizant of this happening. In the long run, over many replications of the experiment, such things can only occur with a very low frequency. So in the long run, randomization gives a firm basis for cause and effect conclusions. But one still needs to be careful in each particular case.

In medicine, specifically in epidemiology, people take another approach to establishing cause and effect. In statistical experiments, the idea is to create a pool of experimental material and then randomly assign treatments to the experimental material. The only systematic differences in the experimental material are the treatment differences, hence treatments cause observed effects. *The epidemiologic approach is to try to adjust for every systematic difference in the treatment groups.* If one properly adjusts for every systematic difference in the treatment groups, and differences between the treatment groups remain after all of this adjustment, the differences must be due to the treatments. This is a reasonable idea. The difficulty is in the execution. How does one appropriately adjust for every (important?) systematic difference in the treatment groups? In a medical example a variety of factors immediately present themselves as candidates for adjustment: height, weight,

Table 9.1: *Summary Statistics, Vinyl Covers*

Vinyl Cover	N	$\bar{y}_{i\cdot}$	s_i^2
A	10	2273.6	16466.5
B	10	2412.8	39867.7
C	10	2222.4	11517.2

blood pressure, cholesterol, etc. But how does one ever know that *all* the *important* factors have been taken into account? That presumes one knows all of the important factors. Moreover, how does one ever know whether the form of these adjustments has been appropriate?

When studying ages of suicides for three racial groups, one might decide to adjust for various socio-economic factors. Were there differences in the three racial groups? Yes. Were there differences in the three racial groups after adjusting for socio-economic factors? Whether "yes" or "no", it is merely the answer to a slightly different question. And in this case, if differences exist, that will still never tell us what it is about the lives (social or genetic) of these groups that cause such differences.

Two primary goals of experimental design are to ***reduce the variability*** *that treatment comparisons are subject to and to provide for a* ***valid estimate of the variability*** *of observations and estimates.* There are several standard experimental design techniques for providing valid variance estimates and reducing variability. The completely randomized design gives a valid variance estimate. *Variance reduction is generally attained by comparing treatments within* ***blocks*** *of relatively homogeneous experimental material.* Randomized complete block designs, Latin Square designs, and incomplete block designs all use forms of blocking to reduce variability and all provide for valid estimates of variances.

Two additional features of experimental design will be discussed. In Section 4 we discuss a very efficient way of defining treatments for an experiment. This is called creating a *factorial treatment structure.* In Chapter 10 we discuss *fractional replications*, which provide methods for obtaining information on the main effects of many factors using relatively few observations.

9.1 One-way Anova

One-Way ANOVA is a method for evaluating whether different groups have the same mean value. It may seem curious that a method for evaluating means would be called "analysis of variance" but we will show why that is.

EXAMPLE 9.1.1 *Vinyl Floor Covering Data*.
Table 2.3 gives data on tear test values for three vinyl floor coverings. In Section 2.4 we used box plots to compare the three coverings. Here we present a more formal approach to evaluating whether the groups have different population mean tear levels. Table 9.1 gives summary statistics for the three covers. The two numbers computed from each group are the sample mean, $\bar{y}_{i\cdot}$, and the sample variance, s_i^2. Due to random variation in the three samples, the sample means for the three covers are all different. Our quest is to find out whether the sample means are so different as to suggest that the groups actually have different population means.

One-way ANOVA is a comparison of two variance estimates. One estimate is valid all of the time, and one estimate is only valid if the groups have the same population mean value. Our underlying assumptions are that all the observations are independent, all the observations have the same variance, say σ^2, and the observations within each group all have the same mean value μ_i. We are testing whether the μ_is are all equal.

From each group we get an estimate of the variance s_i^2. Since the variance is supposed to be the same in each group, averaging the s_i^2s should give a better estimate of σ^2 than any particular s_i^2.

Thus our estimate of σ^2 that is valid all of the time is the average of the s_i^2s. This is called the *mean squared error* and written *MSE*. For the vinyl cover data, this pooled estimate of the variance is

$$MSE = \frac{16466.5 + 39867.7 + 11517.2}{3} = 22617.$$

The other estimate of the variance σ^2 is a bit more complicated. Recall from Chapter 3 that if $y_1, \ldots, y_N$ are a random sample with $\mathrm{E}(y_i) = \mu$ and variance $\mathrm{Var}(y_i) = \sigma^2$, the mean $\bar{y}_{\cdot}$ has expected value $\mathrm{E}(\bar{y}_{\cdot}) = \mu$ and variance $\mathrm{Var}(\bar{y}_{\cdot}) = \sigma^2/N$. The trick is to apply these results to each of the group sample means. If the μ_is are all the same, the $\bar{y}_{i\cdot}$s all have the same expected value and they all have the same variance, σ^2/N. (In the floor covering example $N = 10$.). Moreover, since all the observations are independent, the observations within each group are independent of observations in different groups, so the $\bar{y}_{i\cdot}$s are all independent. That means we can treat the $\bar{y}_{i\cdot}$s as a random sample. If the $\bar{y}_{i\cdot}$s are a random sample, we can compute their sample variance and it will provide an estimate of the variance of an individual $\bar{y}_{i\cdot}$, i.e., σ^2/N. To get an estimate of σ^2, just multiply the sample variance of the $\bar{y}_{i\cdot}$s by N, the number of observations in each group. This is called the *mean squared groups* and written *MSGrps*. For the vinyl cover data, if you plug the $\bar{y}_{i\cdot}$ numbers, 2273.6, 2412.8, and 2222.4, into a hand calculator or computer to get the sample variance, the value is 9708.4 and

$$MSGrps = 10(9708.4) = 97084.$$

This variance estimate has built into it the assumption that the different vinyl covers have no mean differences between them. If that is not the case, *MSGrps* will involve both the variability of an observation, σ^2, and the variability between the groups means, the μ_is. Thus, if the μ_is are not all equal, *MSGrps* tends to be larger than σ^2.

Finally, to evaluate whether the μ_is are different, we compare our two variance estimates. If there are no differences, *MSGrps* and *MSE* should be about the same, i.e., their ratio $F \equiv MSGrps/MSE$ should be close to 1. If the μ_is are different, *MSGrps* tends to be larger than *MSE*, so the ratio $F \equiv MSGrps/MSE$ tends to be larger than 1. If $MSGrps/MSE$ is much larger than 1, we conclude that there must be differences between the μ_is. For the vinyl data, $F = 97084/22617 = 4.29$.

Question: "Is 4.29 much larger than 1?" In order to answer that question, we quantify the random variability of such F values when there really are no differences among the μ_is. If our observed F value is so much larger than 1 that it would rarely be observed when the μ_is are all equal, that suggests that the μ_is are probably not all equal. When the μ_is are all equal (and the data all have normal distributions), the F value is an observation from a distribution called an F distribution with, for the vinyl data, 2 degrees of freedom in the numerator and 27 degrees of freedom in the denominator. One can look up, say, the 95th percentile of an $F(2,27)$ distribution, it is about 3.35. If the μ_is are all equal, 95% of the time we would get an observed F value of 3.35 or less. We actually saw $F = 4.29$, which is a strange thing to happen when the μ_is are all equal, but not necessarily strange when they are different, so we conclude that the μ_is are not all equal.

The degrees of freedom for the F test were computed as follows. The estimate of the variance in the numerator was based on computing the sample variance of the group means. This is just a sample variance, based on 3 observations because there are three groups, so the sample variance has $3-1=2$ degrees of freedom. *MSGrps* also has 2 df. The estimate of the variance in the denominator was based on averaging the sample variances from each group. Each group has $N = 10$ observations, so each group sample variance has $N-1=9$ df. By averaging, we are pooling information from all of the groups, so the mean squared error has degrees of freedom equal to the sum of the df for each group variance, i.e., the degrees of freedom for error are $dfE = 3(9) = 27$, where 9 is the degrees of freedom from each group and 3 is the number of groups.

The ANOVA calculations for the vinyl floor covers are summarized in the ANOVA table, Table 9.2. The sources are Covers (groups), Error, and Total. The entries in the row for Total, where they exist, are just the sum of the entries for Covers and Error. The computations in the df, *MS*, and F columns have already been explained. The *SS* column contains sums of squares for Covers, Error,

Table 9.2: *Analysis of Variance for Vinyl Floor Coverings*

Analysis of Variance					
Source	*df*	*SS*	*MS*	*F*	*P*
Covers	2	194167	97084	4.29	0.024
Error	27	610662	22617		
Total	29	804830			

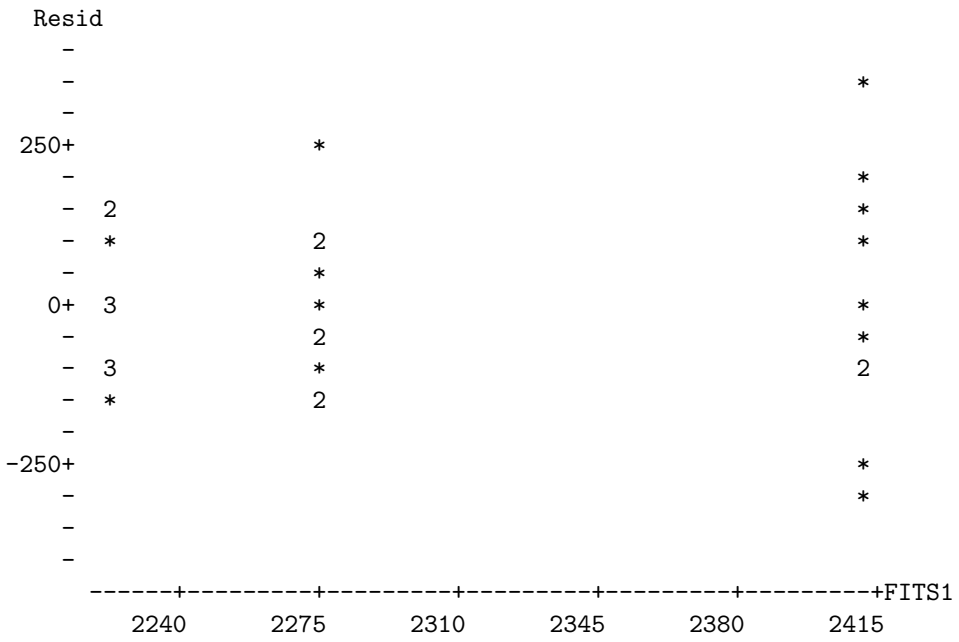

Figure 9.1: *Vinyl Floor Covering: Residuals versus predicted values*

and Total. The first two of these are just the product of the corresponding MS and the df. Finally, the column labeled P is the probability of getting a value larger than the computed F value when the μ_is are all equal. The row for Total can also be computed directly. Take all of the observations from the experiment, regardless of group, and compute the sample variance. The degrees of freedom Total are the degrees of freedom for this estimate, i.e., the total number of observations minus 1. The sum of squares total is just this sample variance of all the observations times the degrees of freedom total.

The ANOVA calculations are based on the assumption that the observations have equal variances and normal distributions. To evaluate these assumptions we compute residuals as the difference between an observation and the sample mean for observations in its group. Thus from Table 9.1, observations from Cover A have 2273.6 subtracted to give the residual and observations from Cover B have 2412.8 subtracted. We can plot the residuals against the group means to get an idea of whether the variabilities change from group to group. Such a plot is given in Figure 9.1. An alternative method for evaluating whether the variabilities are constant over the groups is too compare the s_i^2s. But remember, there is natural variability among the s_i^2s, even when the variances of all observations are the same. There is some hint that the group variance tends to increase as the group means get larger. This is not an uncommon phenomenon. We might reanalyze the data by replacing each observation with its square root or, alternatively, replacing each observation with its logarithm. These might eliminate the problem. However, in the current case, the evidence for increasing variances is not so strong that a corrective measure really seems necessary.

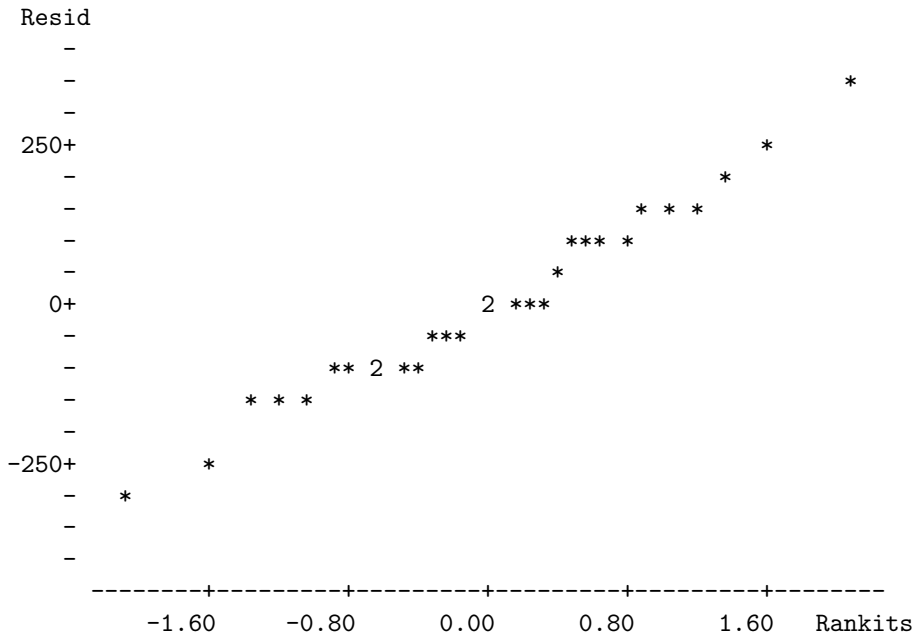

Figure 9.2: *Vinyl Floor Covering: Residuals versus normal scores,* $W' = 0.99^2$

We can also perform a residual plot to evaluate the normality of the data. Figure 9.2 is a normal plot of the residuals. If the data are normal, the normal plot should be approximately straight. This plot looks quite good. □

9.1.1 ANOVA and Means Charts

The data required for a one-way ANOVA and the data required for a means chart have identical structures. In both cases we have various groups and a collection of observations from each group. In typical one-way ANOVA problems the number of groups tends to be rather small and the number of observations in each group somewhat larger. In a typical means chart the number of groups is large and the number of observations in each group is relatively small. But one-way ANOVA can be applied any time there are groups and observations within groups. In this subsection we review the one-way ANOVA procedures while introducing more terminology, mathematical notation, and generalizations. We also relate ANOVA to means charts.

The basic idea of a means chart in statistical process control is to evaluate whether the data appear to be a random sample from some population. If data form a random sample, we can use them to predict future behavior of the process. In one-way ANOVA we exploit a formal statistical test of this idea.

The blood pressure data given in Table 4.2 are repeated in Table 9.3. We have 19 groups of 4 observations. More generally, we will think about having a groups, each with N observations, say y_{ij}, $i = 1,\ldots,a$, $j = 1,\ldots,N$. (In means charts we called the number of groups n because that was the number of points we were plotting. Now we use a.) We want to test the model that the observations are iid which typically includes that they are (a) all independent, (b) all have the same mean value, say, μ, (c) all have the same variance, say, σ^2. Frequently we also assume that (d) they all have normal distributions. Our test focuses on the possibility that observations in different groups have different mean values, say, μ_i for group i. In the context of control charts, the groups are formed from rational subgroups, i.e., observations that are taken under essentially identical conditions, so observations should have the same mean within each group.

Table 9.3: *Nineteen Samples of Systolic Blood Pressures*

Group	Data				N	$\bar{y}_{i\cdot}$	s_i	Range
1	105	92	98	98	4	98.25	5.32	13
2	95	96	84	93	4	92.00	5.48	12
3	90	97	97	90	4	93.50	4.04	7
4	90	88	86	90	4	88.50	1.915	4
5	100	92	92	90	4	93.50	4.43	10
6	78	82	82	84	4	81.50	2.52	6
7	90	83	83	87	4	85.75	3.40	7
8	95	85	88	90	4	89.50	4.20	10
9	92	90	87	85	4	88.50	3.11	7
10	84	85	83	92	4	86.00	4.08	9
11	90	87	84	90	4	87.75	2.87	6
12	93	88	90	88	4	89.75	2.36	5
13	92	94	87	88	4	90.25	3.30	7
14	86	87	91	88	4	88.00	2.16	5
15	90	94	94	82	4	90.00	5.66	12
16	95	95	87	90	4	91.75	3.95	8
17	96	88	90	84	4	89.50	5.00	12
18	91	90	90	88	4	89.75	1.258	3
19	84	82	90	86	4	85.50	3.42	8
Total					76	89.434	4.848	7.947

More mathematically, we assume (a) that the y_{ij}s are all independent, (b) that $\mathrm{E}(y_{ij}) = \mu$, versus the possibility that $\mathrm{E}(y_{ij}) = \mu_i$, (c) that $\mathrm{Var}(y_{ij}) = \sigma^2$, and the least important of the assumptions is (d) $y_{ij} \sim N(\mu_i, \sigma^2)$. Dropping only the normality assumption gives us a model with weaker assumptions that leads us to everything except the exact distribution of the F statistic. The weaker one-way ANOVA model is

$$y_{ij} = \mu_i + \varepsilon_{ij}; \qquad \varepsilon_{ij}\text{s} \quad \text{iid}, \quad \mathrm{E}(\varepsilon_{ij}) = 0, \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma^2,$$

with the F test focusing on the additional assumption [null hypothesis] $\mu_i = \mu$ for all i.

As illustrated in Table 9.3, from each group of observations we compute two summary statistics, the sample mean

$$\bar{y}_{i\cdot} \equiv \frac{y_{i1} + y_{i2} + \cdots + y_{iN}}{N}$$

and the sample variance

$$s_i^2 \equiv \frac{(y_{i1} - \bar{y}_{i\cdot})^2 + (y_{i2} - \bar{y}_{i\cdot})^2 + \cdots + (y_{iN} - \bar{y}_{i\cdot})^2}{N-1}.$$

The $\bar{y}_{i\cdot}$s estimate the mean value μ_i from each group. In particular, $\mathrm{E}(\bar{y}_{i\cdot}) = \mu_i$, $\mathrm{Var}(\bar{y}_{i\cdot}) = \sigma^2/N$, and, if the data are normally distributed,

$$\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/N).$$

The s_i^2s estimate the variance of the observations in the ith group which by assumption is σ^2, i.e., $\mathrm{E}(s_i^2) = \sigma^2$. We are also assuming that *all* of the observations are independent, from which it follows that each group of observations is independent of every other group of observations, which in turn implies that the pairs $(\bar{y}_{1\cdot}, s_1^2), (\bar{y}_{2\cdot}, s_2^2), \ldots, (\bar{y}_{a\cdot}, s_a^2)$ are all independent. In particular the $\bar{y}_{i\cdot}$'s are independent, with the same variance σ^2/N, but possibly different mean values μ_i. These are almost a random sample. In fact, the $\bar{y}_{i\cdot}$s would be a random sample of size a if they had the same mean value and that is exactly what we want to test. Similarly, the s_i^2s are a random sample, so taking their sample mean is appropriate.

We will combine the s_i^2s into a relatively obvious estimate of σ^2 and then we will use the $\bar{y}_{i\cdot}$'s, along with the assumption that $\mu_i = \mu$ for all i, to get a less obvious estimate of σ^2.

Getting right to the chase, compute the mean squared error

$$MSE \equiv s_p^2 \equiv \frac{s_1^2 + s_2^2 + \cdot + s_a^2}{a}. \tag{9.1.1}$$

Alternatively, s_p^2 is sometimes used to indicate that the variance estimates have been "pooled" into one estimate, the subscript "p" indicating that it is a pooled estimate of the variance. For the blood pressure data,

$$MSE \equiv s_p^2 = \frac{275.1}{19} = 14.48.$$

Combining the degrees of freedom for each estimate gives $dfE = a(N-1)$ which for the blood pressure data is $19(4-1) = 57$.

Also compute

$$MSGrps \equiv N s_{\bar{y}}^2, \tag{9.1.2}$$

where $s_{\bar{y}}^2$ is the sample variance computed from the $\bar{y}_{i\cdot}$s, i.e.,

$$s_{\bar{y}}^2 \equiv \frac{(\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2 + (\bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot})^2 + \cdots + (\bar{y}_{a\cdot} - \bar{y}_{\cdot\cdot})^2}{a-1} = \frac{\sum_{i=1}^{a} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{a-1}. \tag{9.1.3}$$

Here $\bar{y}_{\cdot\cdot}$ is the sample mean of the $\bar{y}_{i\cdot}$s. For the blood pressure data it is

$$\bar{y}_{\cdot\cdot} = 89.434$$

and we get

$$s_{\bar{y}}^2 \equiv 13.025,$$

so

$$MSGrps = 4(13.025) = 52.1.$$

MSGrps has the same degrees of freedom as $s_{\bar{y}}^2$, i.e., $dfGrps \equiv a - 1$. For the blood pressure data that is $19 - 1 = 18$ degrees of freedom for the groups.

When the μ_is all equal μ, the $\bar{y}_{i\cdot}$s are independent observations with $\mathrm{E}(\bar{y}_{i\cdot}) = \mu$ and $\mathrm{Var}(\bar{y}_{i\cdot}) = \sigma^2/N$, so $s_{\bar{y}}^2$ estimates σ^2/N, i.e.,

$$\mathrm{E}(s_{\bar{y}}^2) = \sigma^2/N$$

and *MSGrps* estimates σ^2, i.e.,

$$\mathrm{E}(MSGrps) = N(\sigma^2/N) = \sigma^2.$$

For the blood pressure data, the 19 values of $\bar{y}_{i\cdot}$ are each one observation from a population with variance $\sigma^2/4$.

The model is that all observations are independent, that the variances of all observations are the same, and that the observations all have the same mean value. If this model is correct, *MSGrps* and *MSE* should be estimating the same number, σ^2. Their ratio, $F \equiv MSGrps/MSE$, should be about 1. Of course there will be variability associated with the F statistic. In particular, when the data have normal distributions, the F statistic will have an F distribution. Formally, the $\alpha = .05$ level test would be to reject the model if

$$\frac{MSGrps}{MSE} > F(1-\alpha, a-1, a(N-1)).$$

For the blood pressure data $F = 3.60$ and $F(1-\alpha, a-1, a(N-1)) = F(.95, 18, 19(4-1)) = 1.788$, so the model is rejected. The idea is that if $MSGrps/MSE$ is too big to reasonably come from the F distribution determined by the model, then something must be wrong with the model. The data are not iid.

One thing that could be wrong with the model is that $\mu_i \neq \mu$ for all i. We now examine the F statistic when this is true. In particular, this violation of the model assumptions tends to make the F statistic larger than it would be if the assumptions were true. It turns out that in general,

$$\mathrm{E}(MSGrps) = \sigma^2 + Ns_\mu^2$$

where s_μ^2 is the "sample variance" of the unobservable μ_is,

$$s_\mu^2 \equiv \frac{(\mu_1 - \bar{\mu}_\cdot)^2 + \cdots + (\mu_a - \bar{\mu}_\cdot)^2}{a-1}.$$

Of course $s_\mu^2 = 0$ if the means are all equal and $s_\mu^2 > 0$ if they are not. Having different μ_i's does not affect MSE, it still estimates σ^2, i.e.,

$$\mathrm{E}(MSE) = \sigma^2.$$

It follows that the F statistic is an estimate of

$$\frac{\mathrm{E}(MSGrps)}{\mathrm{E}(MSE)} = \frac{\sigma^2 + Ns_\mu^2}{\sigma^2} = 1 + \frac{Ns_\mu^2}{\sigma^2}.$$

Thus if the means are not equal, F tends to be larger than 1. What makes F much larger than 1, so that it is easy to tell that the data are not iid, involves some combination of having large group sample sizes N, a small variance σ^2 (which results from having homogeneous material), and large differences (variability) among the μ_is. In process control we do not want to reject the iid assumption unless we need to, so we have N fairly small, we are on a never ending quest to make σ^2 smaller, and we do not want to adjust the process until the changes in the μ_is are so substantial that we must. (The smaller the variability σ^2, the easier it will be to tell when the process goes off target.) In the next subsection we discuss how lack of independence can also cause large F statistics.

In our model, we assume that the variances are all the same. (Of course, the sample variance chart can be used to validate that assumption. The s and R charts are also germane.) Because the variances are all the same, we can average them to get one overall estimate of the variance. In situations like the blood pressure data in which every group has the same number of observations, the simple average gives a good estimate of the variance. If we had 4 observations on one group and 40 observations on another group, the second group would give us a much better estimate of the common variance σ^2 than the first, so a simple average is no longer appropriate. We now generalize the one-way ANOVA F test to handle unequal sample sizes.

If the different groups have different numbers of observations, say, N_i observations in group i, we need to weight the estimates based on how much information they contain about σ^2. For technical reasons, each sample variance s_i^2 is essentially based on the equivalent of $N_i - 1$ observations, so the weighted average variance estimate becomes

$$MSE \equiv s_p^2 \equiv \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2 + \cdot + (N_a-1)s_a^2}{(N_1-1) + (N_2-1) + \cdot + (N_a-1)}.$$

Note that when the N_is are all the same, i.e., $N_i = N$ for all i, this weighted average is just the same as the simple average (9.1.1).

The ith group has N_i observations, so s_i^2 has $N_i - 1$ degrees of freedom. When pooling the variance estimates, we get to add the degrees of freedom, so MSE has Error degrees of freedom,

$$dfE \equiv (N_1 - 1) + \cdots + (N_a - 1) = (N_1 + \cdots + N_a) - a.$$

Alas, the $MSGrps$ is not nearly as intuitive when the sample sizes are different,

$$MSGrps \equiv \frac{\sum_{i=1}^a N_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{a-1}$$

Table 9.4: *Analysis of Variance for the Blood Pressure Data of Table 9.3*

Analysis of Variance					
Source	*df*	*SS*	*MS*	*F*	*P*
Groups	18	937.4	52.1	3.60	0.000
Error	57	825.3	14.5		
Total	75	1762.7			

where, if we now define $n \equiv N_1 + \cdots + N_a$,

$$\bar{y}_{\cdot\cdot} \equiv \frac{\sum_{i=1}^{a} N_i \bar{y}_{i\cdot}}{n} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{N_i} y_{ij}}{n}.$$

Again, *MSGrps* reduces to the formulae (9.1.2) and (9.1.3) when the group sizes are equal. In particular, the grand mean for equal N_is simplifies to

$$\bar{y}_{\cdot\cdot} = \frac{\bar{y}_{1\cdot} + \bar{y}_{2\cdot} \cdots + \bar{y}_{a\cdot}}{a}.$$

Again, if the model with equal means and normal distributions is correct, the $F = MSGrps/MSE$ statistic is one observation from an F distribution, i.e.,

$$\frac{MSGrps}{MSE} \sim F(a-1, dfE).$$

A formal α level test is to reject the model if the observed F statistic is larger than $(1-\alpha)100\%$ of all observations from the $F(a-1, dfE)$ distribution, i.e., if

$$\frac{MSGrps}{MSE} > F(1-\alpha, a-1, dfE).$$

Even without normal distributions the rationale that F should be about 1 for iid data remains valid.

EXAMPLE 9.1.2 *Blood Pressure Data*
From the summary statistics given in Tables 4.2 and 9.3, the ANOVA computations are made. The analysis of variance table is given in Table 9.4. Although the F statistics of 3.60 is smaller than the vinyl covers F of 4.29, this test more clearly shows differences between the groups because it has a P value of 0.000. This test has 18 and 57 degrees of freedom, so the F distribution is more concentrated near 1 than is the $F(2,27)$ distribution appropriate for the vinyl covers. □

We conclude this subsection with two additional examples of data for means charts that will be revisited in Section 2.

EXAMPLE 9.1.3 *Hopper Data*
The data were presented in Table 4.13 and discussed in Exercise 4.8.16. Summary statistics were given in Table 4.14. The ANOVA table is given in Table 9.1.5. A plot of residuals versus group means is given in Figure 9.1.3. The control charts from Exercise 4.8.15 as well as the F test show clearly that the process is out of control. Something is wrong. Perhaps a fishbone diagram could help sort out possible causes. □

EXAMPLE 9.1.4 *Injection Data*
Exercise 4.8.16 examined the outside diameters of injection molded bottles in which each sample consisted of the bottles made from the four heads of the injection molding machine. The data are in Table 4.15 with summary statistics in Table 4.16 and are plotted in Figure 9.4. The ANOVA table is given in Table 9.6. A residual plot is given in Figure 9.5. Neither the charts from Exercise 4.8.16 nor the F test given here show a problem, unless you quite properly worry about the P value for the F test being too big. P values near 0 indicate strong effects, but P values near 1 also indicate something strange is going on, as we will see in Section 2. □

Table 9.5: *Analysis of Variance for Hopper Data*

Analysis of Variance					
Source	*df*	*SS*	*MS*	*F*	*P*
Days	19	296380	15599	18.25	0.000
Error	40	34196	855		
Total	59	330576			

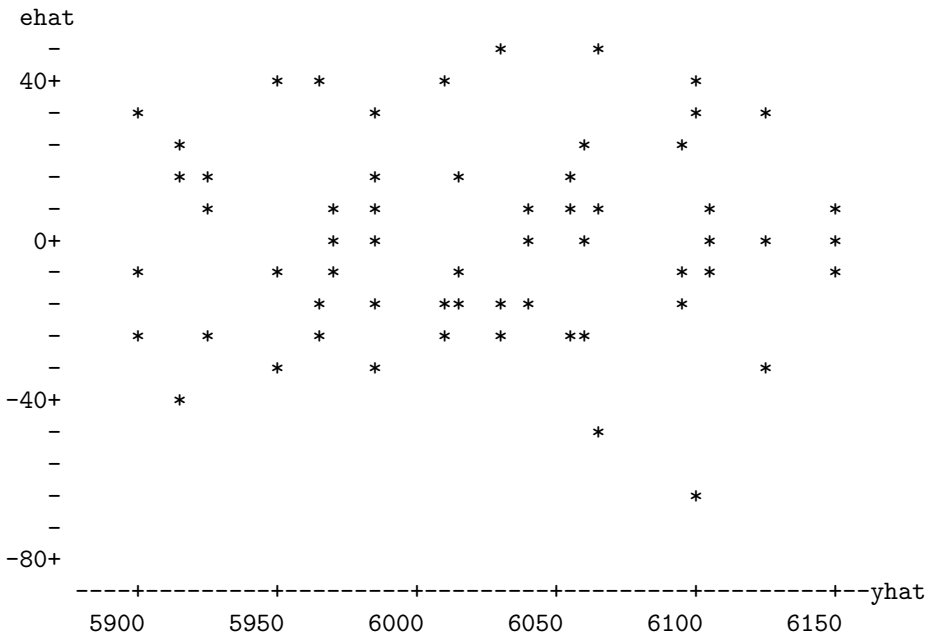

Figure 9.3: *Residual Plot, Hopper data, one factor*

9.1.2 Advanced Topic: ANOVA, Means Charts, and Independence

The most commonly used charts are probably the means charts. They are used to detect changes in the mean of the process. Means charts are constructed based on the assumption that the variance in each group is the same and then use some measure of average variance to determine the width of the control limits. It is not hard to see that if a group has much greater variability than average, it has an increased chance that the group sample mean will exceed the control limits, even if the group population mean is the same as all the other means.

More subtly, means charts are also sensitive to a lack of independence, or at least they are sensitive to positive correlations among the observations within the groups. Center lines and control limits are based on estimates of μ and σ. Without independence, the estimates of μ are still reasonable but lack of independence can cause major problems in the estimate of σ. For example, in the

Table 9.6: *Analysis of Variance: Injection Data*

Analysis of Variance					
Source	DF	SS	MS	F	P
Samples	19	0.021655	0.001140	0.45	0.970
Error	60	0.150300	0.002505		
Total	79	0.171955			

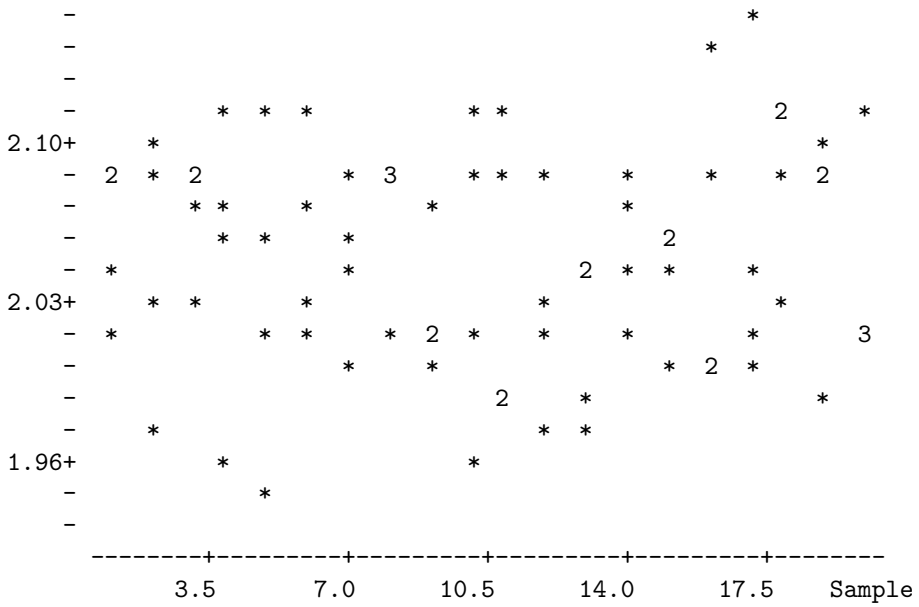

Figure 9.4: *Plot of Injection data*

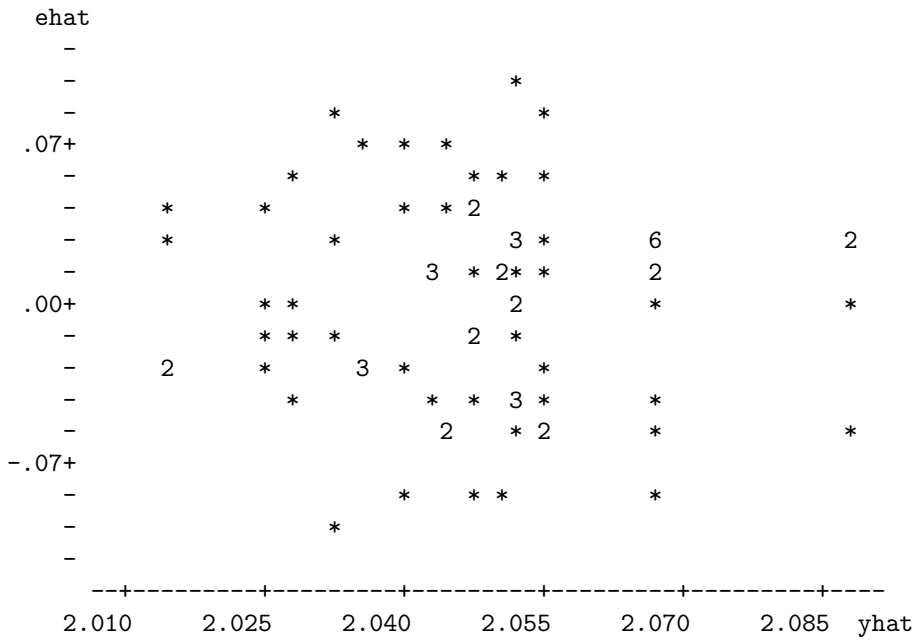

Figure 9.5: *Residual Plot, Injection Data*

means chart, if groups of observations are independent but observations within groups have a constant correlation ρ, then MSE is an unbiased estimate of $\sigma^2(1-\rho)$ rather than of σ^2, see Christensen (1996, p. 240). A similar computation for groups of size N gives $\mathrm{Var}(\bar{y}_{i\cdot}) = (\sigma^2/N)[1+(N-1)\rho]$ rather than σ^2/N. If the correlation is substantially greater than zero, the estimate of σ is much too small while the true variance of the $\bar{y}_{i\cdot}$s is much larger than the nominal value under independence. Both of these features tend to show that the process is out of control. First, the estimated control

limits are inappropriately close to the center line because the estimated control limits approximate $\mu \pm 3(\sigma/\sqrt{N})\sqrt{1-\rho}$ rather than $\mu \pm 3(\sigma/\sqrt{N})$. Second, even if we knew the value of σ, the $\mu \pm 3(\sigma/\sqrt{N})$ control limits are inappropriately close to the center line because the true standard deviation of the $\bar{y}_{i\cdot}$s is $(\sigma/\sqrt{N})\sqrt{1+(N-1)\rho}$ rather than the smaller nominal value of $(\sigma/\sqrt{N})$. Thus, if the observations in the groups have a large positive correlation, we are more likely to obtain observations exceeding the control limits than if the observations were independent. Additionally, many of the other indicators discussed earlier for a process that is out of control are more likely to occur when the data are not independent.

Note that the analysis in the previous paragraph also applies to the one-way ANOVA F test. Even when the population means of the various groups are the same, if the observations within groups have a correlation near 1, the F statistic tends to be large. $MSGrps$ is N times the sample variance of the $\bar{y}_{i\cdot}$s so in this case $MSGrps$ estimates $\sigma^2[1+(N-1)\rho]$; MSE estimates $\sigma^2(1-\rho)$. Thus the F statistic estimates

$$\frac{\sigma^2[1+(N-1)\rho]}{\sigma^2(1-\rho)} = \frac{[1+(N-1)\rho]}{(1-\rho)},$$

which is substantially greater than 1 when ρ is near one. It follows that observed F statistics tend to be large when the observations within groups have a large positive correlation. The F test in Table 9.3 is highly significant so the observations within groups may be correlated or the means of the groups may differ. Without further analysis, one cannot really tell which is the correct conclusion. In either case however, the process is out of control.

Note that the hopper car data of Example 9.1.3 may be a good example of this phenomenon. The ANOVA table showed clear day to day differences. However, it is unlikely that the weight of the car is changing substantially. A more likely explanation is that the three observations made each day are highly correlated with one another.

9.2 Two-Way ANOVA

One-way ANOVA is a method for determining if there are statistically significant differences between groups of observations. Sometimes data are structured in a way that allows us to think of two totally different grouping schemes applied to the same data.

EXAMPLE 9.2.1 *Hopper Data*

Exercise 4.8.16 and Example 9.1.3 discussed data on the weight of a railroad car measured three times on each of 20 days. The data were given in Table 4.13. There are two ways to think about grouping these data. Previously, we thought of each day being a group with 3 observations on each group. However, we could also think in terms of having 3 groups, i.e., the group consisting of the 20 observations that were made first on the various days, another group of second weighings, and a third group of 20 third weighings. With groups being Days, we compute the degrees of freedom and mean square for days just as for a one-way ANOVA, i.e., just as we did in the previous section. Similarly, with the first, second, and third weighings as groups, we can compute 3 sample means, each based on 20 observations. The sample variance of these means times 20 is the mean square for weighings. The total line can be computed as before from the sample variance of all the observations. The error line is obtained by subtraction, i.e., the dfs for Days, Weighings, and Error must add up to the df for Total, and similarly for the SSs. The MSE is SSE/dfE. F tests for Days and Weighings are obtained by dividing their mean squares by the MSE. The ANOVA table is given as Table 9.7. There is no evidence of any systematic differences between the three weighings. As with the means chart which ignored effects due to weighings and the one-way ANOVA that ignored weighing, the two-way ANOVA gives clear evidence of day to day differences. A residual plot for the two-way ANOVA is given as Figure 9.6.

For the hopper data, the order of weighings seemed to have no affect. However, we will see a different story with the injection molding data.

Table 9.7: *Two-Way Analysis of Variance for Hopper Data*

Analysis of Variance					
Source	DF	SS	MS	F	P
Days	19	296380	15599	18.44	0.000
Weighings	2	2049	1024	1.21	0.309
Error	38	32147	846		
Total	59	330576			

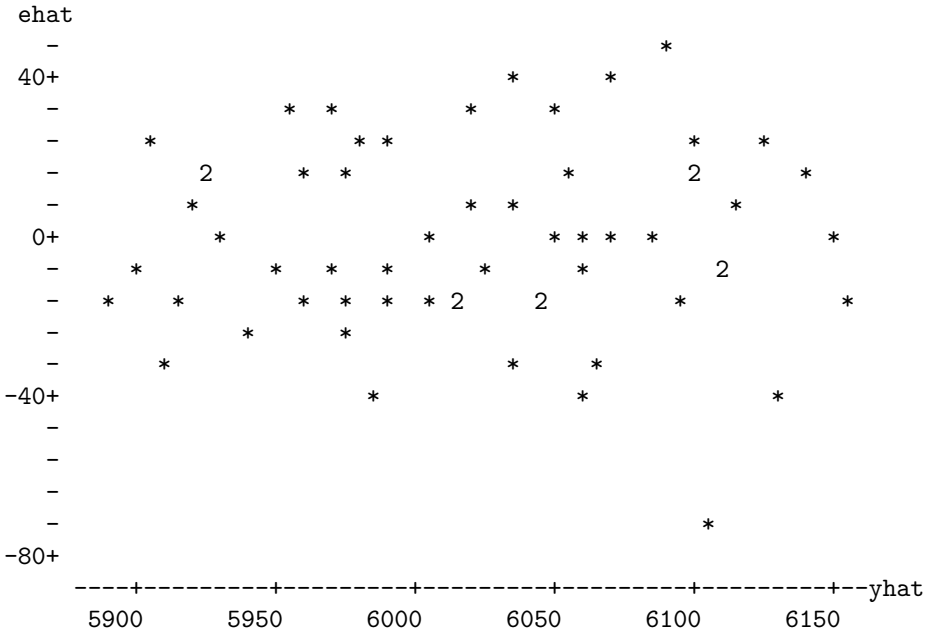

Figure 9.6: *Residual Plot, Hopper data, two factors*

EXAMPLE 9.2.2 *Injection Data*
Exercise 4.8.16 and Example 9.1.4 examined the outside diameters of injection molded bottles in which each sample consisted of the bottles made from the four heads of the injection molding machine. The data are in Table 4.15. As in our previous example, there are two ways to group the data. There are 20 samples each with four observations; the analysis we have used in the past. But we can also think of each head providing a group of 20 observations. The 20 samples provide 19 degrees of freedom for samples and a mean square based on multiplying the variance of the 20 group means by 4. Similarly, the 4 heads provide 3 degrees of freedom for heads and a mean square based on multiplying the sample variance of the 4 head means by the number of observations in each group, 20. The Total line is obtained from the sample variance of all 80 observations. The Error line is obtained by substraction as in the previous example. The ANOVA table is given in Table 9.8. A residual plot is given as Figure 9.7. In Exercise 4.8.16 and Example 9.1.4 in which the heads were ignored, there was no reason to suspect that the process was out of control. However, this analysis gives clear evidence that the four heads are giving bottles with different outside diameters (P value of 0.000). In this case, we originally thought that the four heads were providing rational subgroups — observations that were being taken under essentially identical conditions. However, the two-way ANOVA shows that the heads are making different kinds of bottles, so they do not provide essentially identical conditions.

The P value due to Samples in Table 9.6 was a suspiciously high 0.970. The large P value was

Table 9.8: *Two-Way Analysis of Variance: Injection Data*

Analysis of Variance					
Source	DF	SS	MS	F	P
Samples	19	0.0216550	0.0011397	1.34	0.194
Heads	3	0.1019248	0.0339749	40.03	0.000
Error	57	0.0483750	0.0008487		
Total	79	0.1719548			

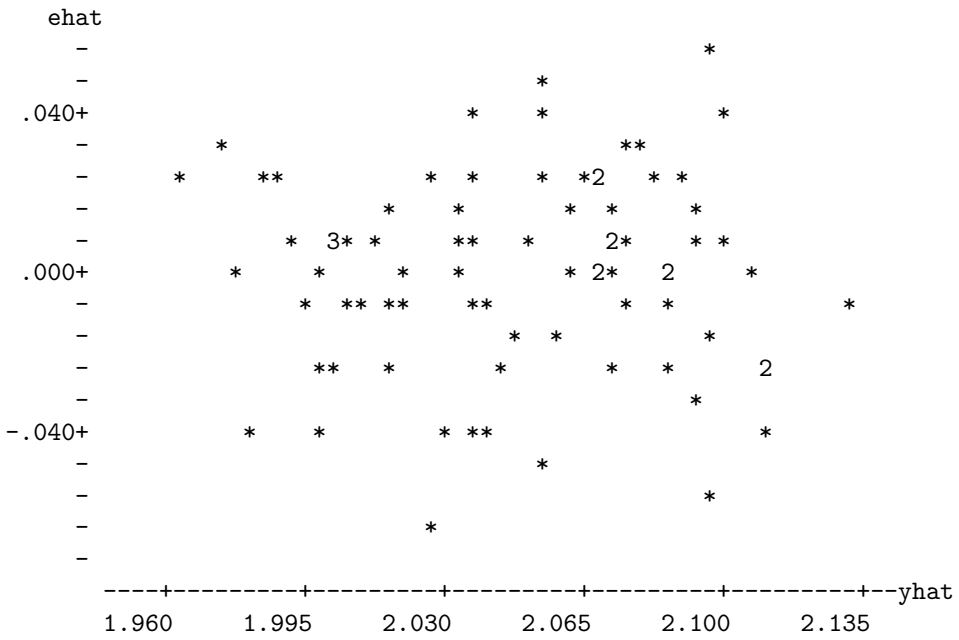

Figure 9.7: *Residual Plot, Injection Data, two factors*

caused by a problem with the iid model. The heads give different results and those differences were incorporated into the Error of Table 9.6. In Table 9.7 we separated the previous Error into two parts, one for Heads and a new Error. The new Error is substantially smaller than the old Error because it no longer incorporates the Head differences. Missing an important factor like that caused the F statistic in Table 9.6 to get unusually small (with an unusually large P value). □

9.3 Basic Designs

9.3.1 Completely randomized designs

The basic idea of a completely randomized design is that a collection of experimental material is prepared and then experimental treatments are randomly assigned to the material. (Equivalently, the material can be randomly assigned to the treatments.) At the beginning of the chapter we discussed an experiment reported by Phillips et al. (1998) that involved tearing pieces of vinyl floor covering. The material to be tested was thoughtfully chosen, carefully prepared, and randomly assigned to different machine operators.

The standard analysis for a completely randomized design is a one-way ANOVA with the treatments defining the groups. In addition to the experiment discussed above that allows evaluation of the four operators, Phillips et al. (1998) also reported the data of Example 9.1.1 on tear measurements for three different types of vinyl floor covering. The background for these data was not

reported as carefully as it was for the other data they report. (The data were not as important to the goal of their article.) But we can presume that it was collected using a completely randomized design, so Example 9.1.1 illustrates the beginning of an analysis for these data. A complete analysis involves carefully evaluating the assumptions made in analyzing the data and adjusting the analysis if any of the assumptions seem invalid. A complete analysis would also not end with establishing that there are differences among the groups. A complete analysis should characterize the differences among the groups. In this case, from looking at the group means it is pretty clear that the primary source of differences between the floor coverings is that Cover B has larger tear values than the other covers. However, statistical methods are often needed to distinguish between patterns in the group means that are real and those that are spurious.

9.3.2 Randomized complete block designs

As discussed in the introduction to this chapter, a key method for reducing the variability associated with comparing experimental treatments is by creating blocks of homogeneous experimental material and assigning the treatments to these blocks. A complete block design involves having blocks that have exactly as many experimental units as there are treatments in the experiment. A *randomized complete block design* randomly assigns the treatments to the units within each complete block. (Any time the blocks are not large enough to accommodate all of the treatments, an *incomplete block design* must be used.)

In the injection molding of Example 9.2.2, if each sample consists of a block of experimental material and the material were randomly assigned to the four heads, the design would be a randomized complete block design.

The analysis for a randomized complete block design uses a two-way ANOVA. The treatments form one set of groups and the blocks form the other set of groups. For the injection data, the treatments are the four heads and blocks are the 20 samples. We saw in the previous section that there are clear differences between heads but no significant differences between the samples. As in the previous subsection, a complete analysis of these data involves carefully evaluating the assumptions made in analyzing the data and adjusting the analysis if any of the assumptions seem invalid. A complete analysis also involves characterizing any differences among the treatment groups. (Differences among the blocks are usually of little interest.)

The fundamental idea of blocking is that experimental material within a block should be homogeneous, i.e., the differences in experimental material should be smaller within blocks than they are between blocks. That is what reduces the error for treatment comparisons. With a randomized complete block design, the differences between blocks are isolated in the analysis, leaving (if blocks were created well) a smaller means squared error than would otherwise be obtained. In Example 9.2.2 we have done a poor job of blocking. The F test for Samples shows that the variability between blocks is not substantially greater than the variability within blocks. So we have not constructed blocks that are going to substantially reduce the variability of our treatment comparisons. In fairness to the original authors, we should point out that while the injection data may form a reasonable approximation to a randomized complete block design, the data collection was not originally set up as a randomized complete block design, so the fact that samples make poor blocks is not surprising.

9.3.3 Latin Squares and Greco-Latin Squares

Latin Squares take the idea of randomized complete blocks a step further. In a randomized complete block, you assign each treatment to some experimental unit in each block. Thus if you pick any block, you observe every treatment within it, but also if you pick any treatment, you have an observation with that treatment from every block. There are two ways of grouping the data: by blocks and by treatments.

In a Latin Square design, there are three ways of grouping the data. These are generally referred

Table 9.9: 3×3 *Latin Square*

		Columns		
		1	2	3
Rows	1	A	B	C
	2	B	C	A
	3	C	A	B

Table 9.10: 3×3 *Graeco-Latin Square*

		Columns		
		1	2	3
Rows	1	$A\alpha$	$B\beta$	$C\gamma$
	2	$B\gamma$	$C\alpha$	$A\beta$
	3	$C\beta$	$A\gamma$	$B\alpha$

to as rows, columns, and treatments. In a Latin Square, the numbers of rows, columns and treatments must be the same. The key aspect to a Latin Square is that for every row, you observe both every treatment and every column. For every column, you observe both every treatment and every row. For every treatment, you also observe both every row and every column.

Table 9.9 shows a Latin Square design with three rows and three columns. The treatments are denoted by Latin letters. Note that each letter appears in each row and in each column.

A Greco-Latin Square involves four ways of grouping the data. These are denoted rows, columns, Latin letters, and Greek letters. Table 9.10 gives a Greco-Latin Square with three rows and three columns. Note that by ignoring the Greek letters you get a Latin Square in the Latin letters. Also by ignoring the Latin letters, you get a Latin Square in the Greek letters. Finally, every Greek letter appears exactly once with each Latin letter, and vice versa.

EXAMPLE 9.3.1 Byrne and Taguchi (1989) and Lucas (1994) give data on a Greco-Latin Square design. The data are values y which measure the force in pounds needed to pull a plastic tube from a connector. Large values are good. While the information is not of immediate relevance for our purposes, the data were collected with a condition time of 24 hours, a condition temperature of 72°F and a condition relative humidity of 25%. There are four grouping factors involved in the design: rows (A), columns (B), Latin letters (C), and Greek Letters (D). The actual grouping factors with their group identifications (levels) are listed below.

A : Interference (a_0 =Low, a_1 =Medium, a_2 =High)

B : Wall Thickness (b_0 =Thin, b_1 =Medium, b_2 =Thick)

C : Ins. Depth (c_0 =Shallow, c_1 =Medium, c_2 =Deep)

D : Percent Adhesive (d_0 =Low, d_1 =Medium, d_2 =High)

The actual data are given in Table 9.11.

A partial analysis of variance table is given in Table 9.12. The ANOVA table includes no F tests because no MSE can be computed from this design. The sums of squares for factors A, B, C, and D use up all of the Total degrees of freedom and sums of squares. This is a problem that we will

Table 9.11: 3×3 *Graeco-Latin Square: Pipe Force Data*

	b_0	b_1	b_2
a_0	c_0d_0 (15.6)	c_1d_1 (18.3)	c_2d_2 (16.4)
a_1	c_1d_2 (15.0)	c_2d_0 (19.7)	c_0d_1 (14.2)
a_2	c_2d_1 (16.3)	c_0d_2 (16.2)	c_1d_0 (16.1)

Table 9.12: *Analysis of Variance for Pipe Force Data*

Analysis of Variance			
Source	*df*	*SS*	*MS*
A	2	12.1756	6.0878
B	2	0.5489	0.2744
C	2	6.8356	3.4178
D	2	2.5156	1.2578
Error	0	0.0000	—
Total	8	22.0756	

Table 9.13: *L_9 Orthogonal Array*

Run	a	b	c	d	Average Thickness
1	0	0	0	0	15.6
4	0	1	1	2	18.3
7	0	2	2	1	16.4
3	1	0	1	1	15.0
2	1	1	2	0	19.7
5	1	2	0	2	14.2
8	2	0	2	2	16.3
6	2	1	0	1	16.2
9	2	2	1	0	16.1

encounter frequently in subsequent chapters. Despite the lack of an Error term and the consequent inability to perform statistical tests, the 4 factors seem to display some substantial differences in their sums of squares.

Again examining the ANOVA table, note that if there were no factor D, so that this was a 3×3 Latin Square with only factors A, B, and C, the ANOVA table line for factor D would actually be an error term. So in a 3×3 Latin Square we get 2 degrees of freedom for error. Two degrees of freedom for error makes for a very bad estimate of variability, but it is better than nothing. More importantly, two degrees of freedom for error suggest that we could run several similar 3×3 Latin Squares and in the analysis pool the estimates of error, and thus obtain a good estimate of variability, cf. Christensen (1996, Secs. 9.3 and 11.4).

In later chapters Greco-Latin Squares will rear their ugly heads in another guise. Table 9.11 can be rewritten as Table 9.13. Note that the treatments and data are exactly the same in the two tables. The method of writing a 3×3 Greco-Latin Square used in Table 9.13 is also referred to as a Taguchi L_9 orthogonal array.

Below is a further breakdown of the ANOVA table. The left side fits a regression model

$$y_{hijk} = \beta_0 + \beta_A h + \beta_B i + \beta_C j + \beta_D k + + \beta_{A2} h^2 + \beta_{B2} i^2 + \beta_{C2} j^2 + \beta_{D2} k^2 + \varepsilon_{hijk}.$$

Notice that adding the two sums of squares for each letter gives the sum of squares presented in Table 9.12. The large sums of squares in Table 9.12 are largely due to one of the two pieces in this table. If anything is going on in C, it is is pretty much a linear trend over the three levels. If anything is going on in A it is largely because the middle level is not behaving like the average of the high and low levels.

Table 9.14: *Aceves-Mijares et al.* 3×3*: Graeco-Latin Square.*

		Columns		
		1	2	3
Rows	1	$A\alpha$ (0)	$B\beta$ (97.9)	$C\gamma$ (112.8)
	2	$B\gamma$ (3)	$C\alpha$ (208.6)	$A\beta$ (104.3)
	3	$C\beta$ (47.9)	$A\gamma$ (55.6)	$B\alpha$ (438.7)
		b_0	b_1	b_2
	a_0	c_0d_0 (0)	c_1d_1 (97.9)	c_2d_2 (112.8)
	a_1	c_1d_2 (3)	c_2d_0 (208.6)	c_0d_1 (104.3)
	a_2	c_2d_1 (47.9)	c_0d_2 (55.6)	c_1d_0 (438.7)

Table 9.15: *Aceves-Mijares et al. (1996):* L_9 *Orthogonal Array.*

Run	A	B	C	D	Average Thickness
1	0	0	0	0	0
4	1	0	1	2	3
7	2	0	2	1	47.9
3	0	2	2	2	112.8
2	0	1	1	1	97.9
5	1	1	2	0	208.6
8	2	1	0	2	55.6
6	1	2	0	1	104.3
9	2	2	1	0	438.7

	Regression on y			
Predictor	Coef	SEQ SS	Predictor	SEQ SS
β_A	4.90000	0.00667	β_A	0.007
β_B	−0.649999	0.48167	β_B	0.482
β_C	1.20000	6.82666	β_C	6.827
β_D	−1.10000	2.40667	β_D	2.407
β_{A2}	−2.46667	12.16889	β_{AB}	0.090
β_{B2}	0.183333	0.06722	β_{BC}	8.670
β_{C2}	−0.0666664	0.00889	β_{CD}	1.562
β_{D2}	0.233334	0.10889	β_{DA}	2.032
Constant (β_0)	15.6000			

The right side of the tabulation fits a model

$$y_{hijk} = \beta_0 + \beta_A h + \beta_B i + \beta_C j + \beta_D k + + \beta_{AB} hi + \beta_{BC} ij + \beta_{CD} jk + \beta_{DA} ki + \varepsilon_{hijk}.$$

For these data, it does not tell as clear a story but note that the sums of squares for A^2, B^2, C^2, and D^2 sum to the same total as the sum of squares for AB, BC, CD, and DA and that the results for the constant, A, B, C, and$D$ are identical (up to roundoff).

9.3.3.1 Additional Example of a Graeco-Latin Square

Aceves-Mijares et al. (1996) give the data in Tables 9.14 and 9.15. The tables provides three alternative methods of denoting the treatments. Aceves-Mijares et al. make no distinction between α and γ, or if you prefer, between d_0 and d_2.

9.4 Factorial treatment structures

Often, more than one factor is of interest in an experiment. For example, when making a fabric, the wear can be effected by both the surface treatment of the fabric and the type of fill used. When

Table 9.16: *Surface finish from a lathe.*

Speed	Feed	Finish	Speed	Feed	Finish
−1	−1	7	−1	−1	9
−1	0	77	−1	0	77
−1	1	193	−1	1	190
0	−1	7	0	−1	9
0	0	75	0	0	80
0	1	191	0	1	191
1	−1	9	1	−1	18
1	0	79	1	0	80
1	1	192	1	1	190

Table 9.17: *One-Way Analysis of Variance for Surface Finish Data.*

Analysis of Variance				
Source	*df*	*SS*	*MS*	*F*
Treatments	8	100718	1799	223
Error	9	64	7	
Total	17	100782		

evaluating waterproofing of a fabric, both the laundry washing the fabric and the laboratory testing the fabric may effect results. In evaluating the results of a spectrometer, both the status of the window and the status of the boron tip may be important. When more than one factor is of interest in an experiment, there is an effective method for defining treatments, i.e., make the treatments all combinations of the factors involved. For example, for a spectrometer, the boron tip may be new or used and the window may be clean or soiled. The factorial treatment structure involves 4 treatments obtained by crossing the two levels of the tip with the two levels of the window. Thus, the treatments are: New-clean, New-soiled, Used-clean, Used-soiled. This is referred to as a 2×2 factorial treatment structure, where $2 \times 2 = 4$ is the number of treatments.

There are two advantages to factorial treatment structures as opposed to running experiments for each factor individually. First, running individual experiments may be misleading. If the effect of the boron tip changes depending on whether the window is clean or soiled, you cannot find that out by running an experiment only on the boron tip. Second, if the effect of the tip does not depend on the window status (and conversely, the effect of the window status does not depend on the boron tip), then one experiment using a factorial treatment structure gives as much information on both the tip and the window as does an experiment on the windows *and* a separate experiment on the boron tips.

EXAMPLE 9.4.1 3×3 *Factorial.*
Collins and Collins (1994) report data on the surface finish of a bar that has been lathed. The surface finish is measured as the distance a probe moved vertically as it was drawn across the bar horizontally. Higher readings indicate rougher surfaces. We can think of this as a completely randomized design with 9 treatments, however the 9 treatments are defined as all combinations of two factors each having three levels. The first factor is the speed of the lathe with three levels: 2500, 3500, or 4500 rpm. The second factor is the rate of feed for the tool with the three levels 0.001, 0.005, and 0.009 inches per revolution of the lathe. The data are given in Table 9.16. The one-way ANOVA table is given as Table 9.17. The F statistic is huge, so there are some differences among the 9 treatments.

As discussed earlier in this chapter, examining the ANOVA table is really only the beginning of a data analysis. A complete analysis involves evaluating assumptions and comparing the treatments. While a detailed examination of such issues is beyond the scope of this book, we now take an intermediate step – we break the line in the ANOVA table for treatments up into components for

Table 9.18: *Factorial Analysis of Variance for Surface Finish Data.*

Analysis of Variance					
Source	DF	SS	MS	F	P
speed	2	25	13	1.76	0.227
feed	2	100670	50335	7078.38	0.000
speed*feed	4	23	6	0.80	0.556
Error	9	64	7		
Total	17	100782			

the different factors in the treatment structure. In this case, we have lines for speeds, the feeds, and interactions between the speeds and feeds. Interaction, evaluates whether the effects of speeds changes depending on what feed is being used, or equivalently, it evaluates whether the effects of feeds changes depending on what speed is being used. Table 9.18 gives the expanded ANOVA table.

To compute the expanded ANOVA table, first ignore the feeds altogether. Pretend that this is an experiment with only three treatments: the three speeds. Compute group means for the three speeds, and perform the usual one-way ANOVA computation, i.e., find the sample variance of the three speed means and multiply by the number of observations that went into each speed mean. This process gives the mean square for speeds with $3-1=2$ degrees of freedom and a sum of squares obtained by multiplying the the degrees of freedom and the mean square. A similar computation gives the df, MS, and SS for feeds. Finally, the interaction is that part of the Treatments from Table 9.17 that is left over after isolating the speed and feed effects. In other words, the interaction sum of squares is obtained by subtracting the speed sum of squares and the feed sum of squares from the sum of squares for treatments. The interaction degrees of freedom are found by a similar subtraction. F tests in Table 9.18 are computed as the ratio of the means squares for speed, feed, and interaction divided by the MSE. In evaluating Table 9.18, the first order of business is to examine the interaction. A significant interaction means that the effects of speeds change depending on what feed is used (and that the effect of feeds depend on what speed is used). If there is interaction, there must be effects for both speeds and feeds. If the effects of speeds change depending on what feed is used then there must be some effect for speeds and if they can change with feeds, the feeds are obviously affecting the results also.

The line in the ANOVA table for speeds examines whether there are speed effects when averaging over the feeds used in the experiment. If there is no interaction, the speed effects do not depend on feeds, so this just measures the speed effects. If there is interaction, there really is nothing we can call speed effects. When there is interaction, speed effects change depending on feeds. Therefore, when there is interaction, looking at speed affects by averaging over the feeds is typically *not* a very useful thing to do. For these data, there is no evidence of interaction so looking as speed effects is reasonable. Similarly, looking at feed effects is only reasonable when there is no interaction. From Table 9.18, we see no effect for speeds but a substantial effect for feeds. It would be appropriate now to examine specific comparisons between the results of using the different feeds.

Finally, we should evaluate the assumptions. Figures 9.8 and 9.9 give residual plots for the analysis. The normal plot looks fine, but the residual plot versus predicted values suggests that the variances are decreasing with the size of the predicted values. Table 9.19 contains the data, residuals, and fitted values. □

To conclude we examine a much more challenging experimental design that involves 4 factors each at 2 levels.

EXAMPLE 9.4.2 2^4 *Factorial.*
Collins and Collins (1994, qe94-559) also looked at surface finish in another experiment with four factors each at two levels. This involved a computer controlled lathe that is fed a bar of stock a certain number of inches per revolution while a tool cuts it. A collet holds the stock in place. The

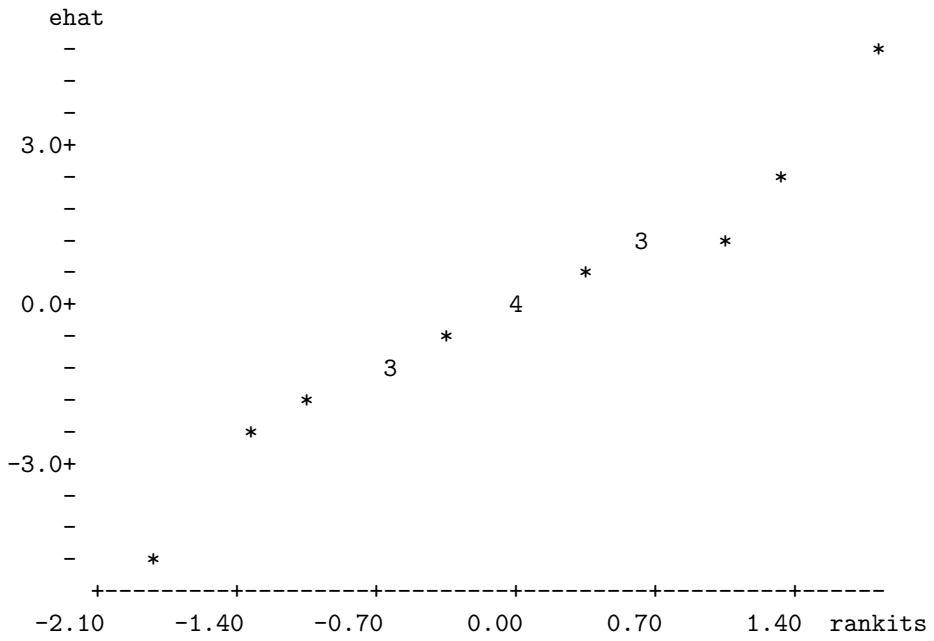

Figure 9.8: *Normal Plot of Surface Finish Data*

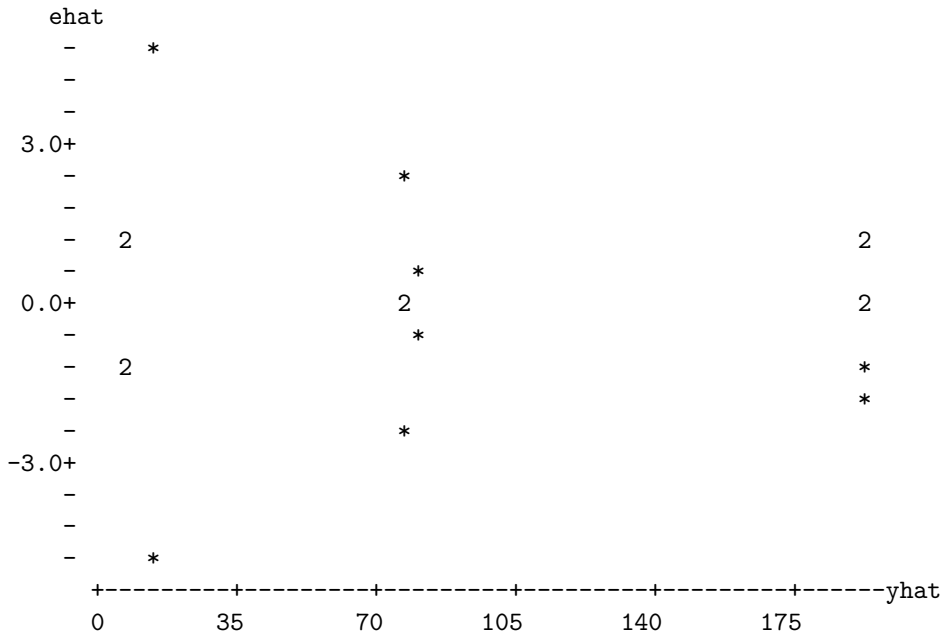

Figure 9.9: *Residual Plot of Surface Finish Data*

four factors and their levels are given in Table 9.20. The data involve two replications of the 16 treatments and are given in Table 9.21. Examining the data, there is little need for a formal analysis. y measures roughness and all of the smallest measures are associated with having Feed at 0.003 and a tight Collet. Nonetheless we will use this example to illustrate some tools for analysis.

An ANOVA conducted on the raw data resulted in the residuals versus predicted values plot of

Table 9.19: *Residuals and Fitted Values: Surface Finish Data*

Case	Speed	Feed	y	$\hat{e}$	$\hat{y}$
1	−1	−1	7	−1.0	8.0
2	−1	0	77	0.0	77.0
3	−1	1	193	1.5	191.5
4	0	−1	7	−1.0	8.0
5	0	0	75	−2.5	77.5
6	0	1	191	0.0	191.0
7	1	−1	9	−4.5	13.5
8	1	0	79	−0.5	79.5
9	1	1	192	1.0	191.0
10	−1	−1	9	1.0	8.0
11	−1	0	77	0.0	77.0
12	−1	1	190	−1.5	191.5
13	0	−1	9	1.0	8.0
14	0	0	80	2.5	77.5
15	0	1	191	0.0	191.0
16	1	−1	18	4.5	13.5
17	1	0	80	0.5	79.5
18	1	1	190	−1.0	191.0

Table 9.20: *Surface Finish Factors*

	Levels		
Factor	−1	1	Units
Speed	2500	4500	rpm
Feed	0.003	0.009	in./rev.
Collet	Loose	Tight	
Tool Wear	New	After 250 parts	

Figure 9.10. The plot looks like evidence that the variability increases with the mean values, although the impression is almost entirely due to the large discrepancy between the two observations 273 and 691 corresponding to high speed, low feed, loose collet, and a new tool. As a corrective measure we analyzed the reciprocals of the square roots of the data, i.e., $1/\sqrt{y}$. Since y is a measure of roughness, $1/\sqrt{y}$ is a measure of smoothness. A one-way ANOVA for $2^4 = 16$ treatments on the transformed data results in the much more attractive residual – predicted value plot given

Table 9.21: *Surface Finish*

Speed	Feed	Collet	Wear	Finish	Speed	Feed	Collet	Wear	Finish
1	1	1	1	216	1	1	1	1	217
−1	1	1	1	212	−1	1	1	1	221
1	−1	1	1	48	1	−1	1	1	39
−1	−1	1	1	40	−1	−1	1	1	31
1	1	−1	1	232	1	1	−1	1	235
−1	1	−1	1	248	−1	1	−1	1	238
1	−1	−1	1	514	1	−1	−1	1	437
−1	−1	−1	1	298	−1	−1	−1	1	87
1	1	1	−1	238	1	1	1	−1	245
−1	1	1	−1	219	−1	1	1	−1	226
1	−1	1	−1	40	1	−1	1	−1	51
−1	−1	1	−1	33	−1	−1	1	−1	33
1	1	−1	−1	230	1	1	−1	−1	226
−1	1	−1	−1	253	−1	1	−1	−1	214
1	−1	−1	−1	273	1	−1	−1	−1	691
−1	−1	−1	−1	101	−1	−1	−1	−1	130

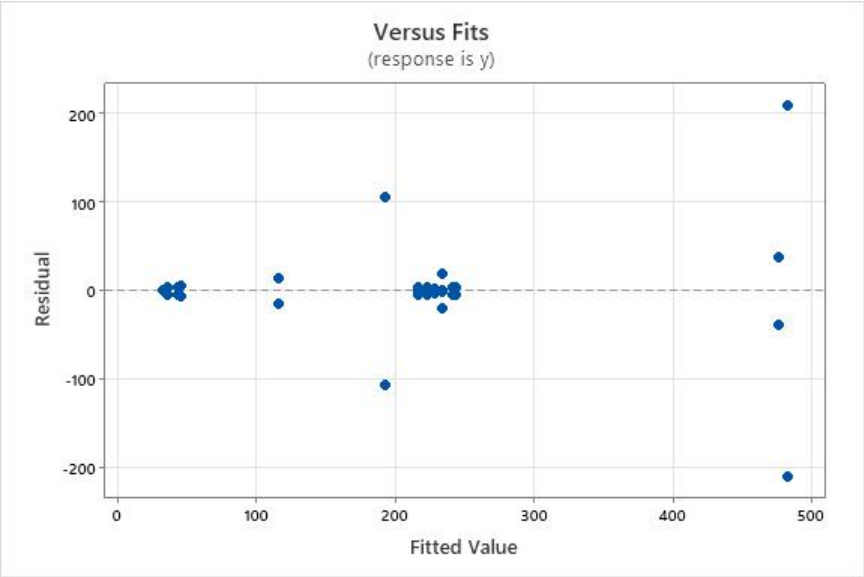

Figure 9.10: *Residual Plot for ANOVA on y.*

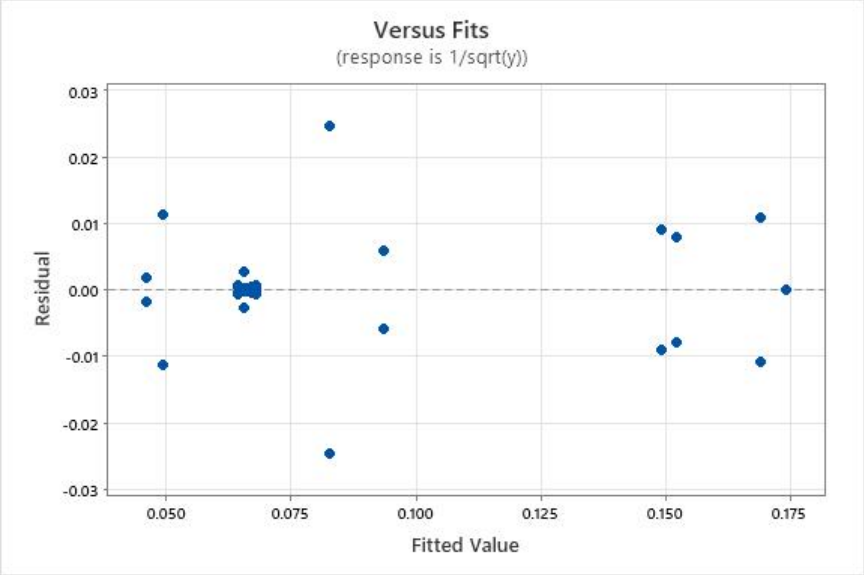

Figure 9.11: *Residual Plot for ANOVA on* $1/sqrt(y)$.

in Figure 9.11. (We tried several transformations and $1/\sqrt{y}$ had the nicest residual plot.) The corresponding normal plot in Figure 9.12 looks tolerable. It includes too many residuals too close to 0.

In the ANOVA of Table 9.22, the 15 degrees of freedom for the 16 treatments have been divided into 15 terms. Four are for the average effects of Speed, Feed, Collet, and (Tool) Wear. Six are for two factor interactions. For example the Speed*Feed interaction looks at whether the average effect of Speed changed depending on what Feed we are looking at. Another four terms are three factor interactions. For example the Speed*Feed*Collet interaction looks at whether the Speed*Feed

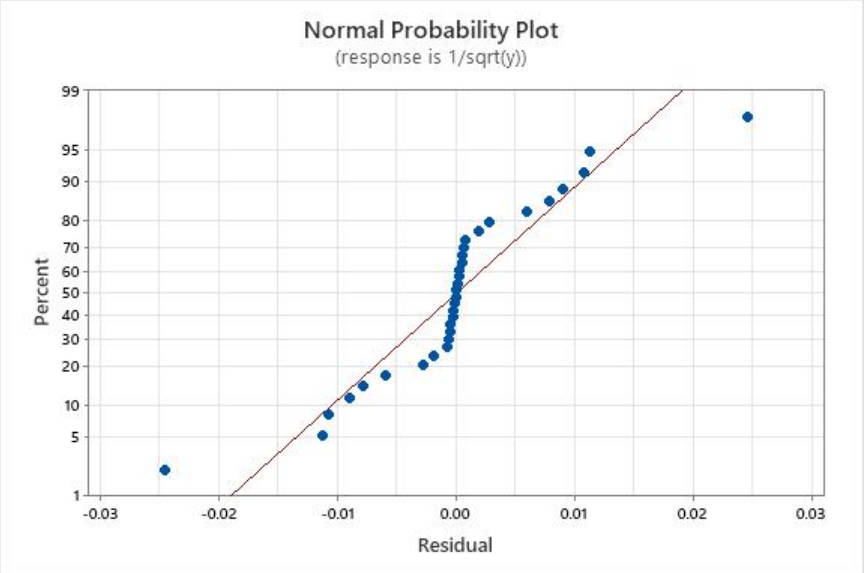

Figure 9.12: *Normal Plot for ANOVA on* $1/sqrt(y)$.

Table 9.22: *Analysis of Variance for* $1/\sqrt{y}$.

Source	DF	SS	MS	F	P
Speed	1	.0019022	.0019022	14.63	.001
Feed	1	.0187102	.0187102	143.86	.000
Collet	1	.0179270	.0179270	137.83	.000
Wear	1	.0000249	.0000249	0.19	.668
Speed*Feed	1	.0018528	.0018528	14.25	.002
Speed*Collet	1	.0001504	.0001504	1.16	.298
Speed*Wear	1	.0000473	.0000473	0.36	.555
Feed*Collet	1	.0168288	.0168288	129.39	.000
Feed*Wear	1	.0000437	.0000437	0.34	.570
Collet*Wear	1	.0000454	.0000454	0.35	.563
Speed*Feed*Collet	1	.0002408	.0002408	1.85	.193
Speed*Feed*Wear	1	.0000203	.0000203	0.16	.698
Speed*Collet*Wear	1	.0000009	.0000009	0.01	.935
Feed*Collet*Wear	1	.0000038	.0000038	0.03	.866
Speed*Feed*Collet*Wear	1	.0000002	.0000002	0.00	.967
Error	16	.0020810	.0001301		
Total	31	.0598798			

interaction changes between a loose and a tight Collet. Finally, the Speed*Feed*Collet*Wear interaction looks at whether the Speed*Feed*Collet changes between having a new tool and having a used tool. All of these interaction descriptions remain valid when changing the factors. For example, Speed*Feed is also how the Feed effect changes with Speed, the Speed*Feed*Collet interaction is also how the Speed*Collet interaction changes depending on the Feed rate, etc. If the four-factor interaction is important, it means that the three factor interactions must exist and be changing, so there is no point in looking at average three-factor interaction. If the Speed*Feed*Collet interaction is important then all of the Speed*Feed, Speed*Collet, and Feed*Collet interactions must exist and be changing. We always investigate the largest interactions first and work our way down. (Down to lower order interactions but physically up the table.)

In these data the primary interaction that looks important is the Feed*Collet interaction but the

Speed*Feed interaction also looks significant. With the important interactions being exactly 2 two-factor interactions that share one common term, the simplest method for interpreting the results is to fix the level of the common factor Feed and look at the separate effects of Speed and Collet for each level of Feed. With no missing data we need only look at the Feed-Collet means and the Feed-Speed means.

	Collet			Speed	
Feed	Loose	Tight	Feed	2500	4500
0.003	0.06786	0.16106	0.003	0.12978	0.09914
0.009	0.06536	0.06683	0.009	0.06620	0.06600

Looking at the Feed-Collet means, it jumps out that the smoothness scores are highest for Feed 0.003 and a tight Collet. The other three scores are all comparable so this is the source of the huge Feed*Collet interaction. Somewhat less obvious is that we can do better at low Speed. From the Feed-Speed means there is little Speed effect at the high Feed rate, but for low Feed, smoothness is higher at Speed 2500 than at Speed 4500. So our formal analysis found us a worthwhile conclusion about Speed that was not obvious from the raw data. Although if you go back to the raw data you can see that among the 8 roughness scores with low Feed and tight Collet, the low Speed scores are always a little lower than the high Speed scores.

The only effect not involved in the important interactions is Wear and Wear has no obvious effect. □

9.5 Exercises

Reanalyze all of the rational subgroup data from the Chapter 4 exercises using one-way ANOVA.

EXERCISE 9.5.1. Reanalyze the hopper car data of Table 4.14 as a two-way ANOVA.

EXERCISE 9.5.2. Reanalyze the injection molding data of Table 4.16 as a two-way ANOVA.

EXERCISE 9.5.3. Schneider and Pruett (1994, qe94-348.dat) look at carbon black DBP values over 12 days with 6 readings per day at 3, 7, and 11, both AM and PM (days start at 3PM). The data are in Table 6.3. Do a time series analysis. Look for day effects and for time within day effects.

EXERCISE 9.5.4. Watch Stuart Hunter's introduction to Experimental Design: Part 1 and Part 2.

Chapter 10

2^n factorials

10.1 Introduction

Often in industrial experimentation there are many factors that need to be explored, accompanied by high costs for obtaining data and limited budgets. When many factors are involved, the number of treatments becomes large, so it may be impractical to obtain data on all combinations of the factors. Even worse, with many treatments it will be impractical to obtain appropriate replications on the treatments so that a truly valid estimate of the error can be obtained. In such cases, we need methods for defining appropriate subsets of the treatments so that observations on these treatment subsets will provide the information necessary to proceed.

We begin with a number of examples of such problems.

EXAMPLE 10.1.1. Taguchi (1987, qe94-319) and Box and Bisgaard (1994) discuss a study to find factors that increase the viscosity of sodium silicate. Nine factors were considered.

		Level	
Label	Factor	0	1
A	Silica sand type	#5	#6
B	Reaction time	4 hr	5 hr
C	Agitation	100	150
D	Reaction temperature	$190 \pm 3^\circ$C	$180 \pm 3^\circ$C
E	Rate of temperature increase	$30 - 35^\circ$C/min	$15 - 20^\circ$C/min
F	Relative excess of silica sand	20%	10%
G	Active carbon type	T_3	3A
H	Quantity of active carbon	0.13%	0.06%
J	Evacuated water temperature	$80 \pm 5^\circ$C	$40 \pm 5^\circ$C

A factorial experiment would involve $2^9 = 512$ treatments. This is a completely unwieldy number. Instead, it was decided to examine only 16 treatments. With 16 observations, and thus 16 degrees of freedom, 1 df is associated with the grand mean (intercept) and 9 are associated with the main effects. That leaves 6 degrees of freedom for doing other things. If there is no interaction, the 6 degrees of freedom would constitute error. □

EXAMPLE 10.1.2. Shina (1991) and Lawson and Helps (1996) [q96-467.dat] consider a study on eight factors that affect a wave-soldering process for printed circuit boards. The factors are A: Preheat Temperature; B: Wave Height; C: Wave Temperature; D: Conv. Angle; E: Flux; F: Direction; G: Wave width. Each factor had two levels being considered so a full factorial involves $2^8 = 256$ treatments. Instead, only 32 treatments were examined. □

EXAMPLE 10.1.3. Sarkar (1997) [qe97-529.dat] examined cleaning efficiencies. The experiment involved five factors: feed rate, setting angle, setting distance, feed setting, and opener setting. A full factorial would involve $2^5 = 32$ treatments. Only 16 treatments were actually run. □

EXAMPLE 10.1.4. Ferrer and Romero (1995) [qe95-750.dat] examined an experiment designed

Table 10.1: *Hare's* $1/2$ *rep. from a* 2^5

Batch	Treatment	Batch	Treatment
1	$a_0b_0c_0d_1e_1$	9	$a_0b_1c_1d_1e_1$
2	$a_1b_0c_1d_1e_1$	10	$a_1b_1c_0d_1e_1$
3	$a_1b_1c_0d_0e_0$	11	$a_0b_0c_1d_1e_0$
4	$a_1b_0c_1d_0e_0$	12	$a_0b_0c_0d_0e_0$
5	$a_0b_1c_0d_0e_1$	13	$a_1b_0c_0d_0e_1$
6	$a_0b_0c_1d_0e_1$	14	$a_1b_1c_1d_0e_1$
7	$a_0b_1c_0d_1e_0$	15	$a_1b_0c_0d_1e_0$
8	$a_1b_1c_1d_1e_0$	16	$a_0b_1c_1d_0e_0$

to improve the adhesive force between polyurethane sheets used in inner linings of various equipment. The factors involved are A: amount of glue; B: predrying temperature; C: tunnel temperature; D: pressure. Reducing the amount of glue or tunnel temperature can save the manufacturer big bucks. In this case the full factorial of $2^4 = 16$ treatments were run, but no replication was performed. □

Our primary example follows.

EXAMPLE 10.1.5. Hare (1988) examined data on excessive variability in the taste of a dry soup mix. The source of variability was identified as the 'intermix', a component of the overall mixture that contained flavorful ingredients such as vegetable oil and salt. Five factors were thought to be of potential importance. Factor A was the number of ports for adding vegetable oil to the large mixer in which intermix is made. This could be either 1 (a_0) or 3 (a_1). Factor B was the temperature at which the mixture was produced. The mixer could be cooled by circulating water through the mixer jacket (b_0) or the mixer could be used at room temperature (b_1). Factor C was the mixing time: either 60 seconds (c_0) or 80 seconds (c_1). Factor D was the size of the intermix batch: 1500 pounds (d_0) or 2000 pounds (d_1). Factor E was the delay between making the intermix and using it in the final soup mix. The delay could be either 1 day (e_0) or 7 days (e_1). With 5 factors each at two levels, there are $2^5 = 32$ different treatments. For example, one of the treatments is the combination of factor levels $a_0b_1c_0d_1e_0$, i.e., one port, mixing at room temperature, 60 seconds of mixing, 2000 pounds of mix, and using the intermix immediately. In fact, these happen to be the standard operating conditions for making intermix. Because the experiment is concerned with variability, several observations were needed for each treatment so that the variance could be estimated reasonably. As a result, it was decided that only 16 of the possible 32 treatments could be considered. Table 10.1 contains a list of the treatments that were actually run. (Standard operating conditions were run in batch 7. The treatments were randomly assigned to the batches.) □

When only 16 of 32 possible treatments are examined, it is referred to as a $1/2$ replicate of a 2^5 design. ($2^5 \times (1/2) = 16$ treatments.) If only 8 of the 32 treatments are examined, the design would be a $1/4$ replicate. ($2^5 \times (1/4) = 8$ treatments.) Sometimes these are referred to as 2^{5-1} and 2^{5-2} designs, respectively.

In this chapter we present a brief introduction to methods for defining and analyzing appropriate fractional replications of experiments in which there are f factors each at 2 levels, i.e., 2^f experiments. Most of our efforts will be devoted to examining fractional replications that involve $n = 2^{f-r}$ observations for some r. Frequently in a 2^f experiment, we will be interested only in estimating the f main effects. To do this one could get by with as little as $f + 1$ observations. The number of observations is the number of degrees of freedom available. One degree of freedom is always associated with the grand mean (intercept), and there is one degree of freedom necessary for estimating each of the main effects. Plackett and Burman (1946) presented designs based on Hadamard matrices that allow efficient estimation of main effects. For a Placett-Burman/Hadamard design to exist,

Table 10.2: *Two groups obtained from a* 2^5 *based on ABCDE*

Even	Odd
$a_0b_0c_0d_0e_0$	$a_0b_0c_0d_0e_1$
$a_0b_0c_0d_1e_1$	$a_0b_0c_0d_1e_0$
$a_0b_0c_1d_0e_1$	$a_0b_0c_1d_0e_0$
$a_0b_0c_1d_1e_0$	$a_0b_0c_1d_1e_1$
$a_0b_1c_0d_0e_1$	$a_0b_1c_0d_0e_0$
$a_0b_1c_0d_1e_0$	$a_0b_1c_0d_1e_1$
$a_0b_1c_1d_0e_0$	$a_0b_1c_1d_0e_1$
$a_0b_1c_1d_1e_1$	$a_0b_1c_1d_1e_0$
$a_1b_0c_0d_0e_1$	$a_1b_0c_0d_0e_0$
$a_1b_0c_0d_1e_0$	$a_1b_0c_0d_1e_1$
$a_1b_0c_1d_0e_0$	$a_1b_0c_1d_0e_1$
$a_1b_0c_1d_1e_1$	$a_1b_0c_1d_1e_0$
$a_1b_1c_0d_0e_0$	$a_1b_1c_0d_0e_1$
$a_1b_1c_0d_1e_1$	$a_1b_1c_0d_1e_0$
$a_1b_1c_1d_0e_1$	$a_1b_1c_1d_0e_0$
$a_1b_1c_1d_1e_0$	$a_1b_1c_1d_1e_1$

the number of observations must be a multiple of 4 but does not need to be a power of 2. Most Plackett-Burman designs are not very good for exploring interactions whereas 2^{f-r} designs make it relatively easy to determine what interaction information is available. We will spend considerable time exploring interactions. A more detailed discussion of these topics is given in Christensen (2017) which includes a modified version of material available in Christensen (1996, Chapter 17).

10.2 Fractional replication

The essence of fractional replication is dividing the 2^f different treatments into equal sized subgroups (whose sizes are also powers of 2). For a $1/2$ replication, we need two groups. The $1/2$ replication involves obtaining data on only one of the two groups. For a $1/4$ replication, we need four groups. The $1/4$ replication involves obtaining data on only one of the four groups. For a $1/8$ replication, we need eight groups and obtain data on only one of the eight groups. After dividing the treatments into groups, it does not matter at all which of the groups you actually use, although some authors like Taguchi, for simplicity, have publicized designs based on one particular group.

Of course if we are only looking at half the treatments, we cannot expect to get as much information as we would if we looked at all of the treatments. Some of the information is going to be lost by using fractional replication. We need to keep track of what things we can and cannot learn from a particular fractional replication. It turns out that many of the effects that we would normally look at in an analysis of variance are "aliased" with each other. When effects are aliased, you cannot tell them apart. They are two names for the same thing. So, for example, we might estimate the main effect for some factor but in a fractional replication it may be impossible to distinguish between that main effect and some three factor interaction effect. If we determine that this effect is important, we have no statistical way of determining if it is the main effect that is important or the three factor interaction that is important. In such a case, the fractional replication will only be useful if we are prepared to **assume** that the three factor effect cannot be important.

EXAMPLE 10.2.1. Hare (1988) used the standard method for breaking a 2^5 into two groups. The two groups are given in Table 10.2. Hare used the group on the left. The treatments are listed in a different order than in Table 10.1 because in Table 10.1 the treatments are listed in the random order in which they were actually run.

At this point the simplest way to tell the two groups in Table 10.2 apart is simply by adding all the subscripts for the treatments. In other words, for the treatment $a_0b_1c_0d_1e_0$ we add the numbers

$0+1+0+1+0=2$. For all the treatments on the left hand side, the sum is even. For all the treatments on the right hand side the sum is odd. □

Different ways of breaking treatments into groups correspond to different patterns of summing subscripts. The different patterns of summing the subscripts in turn determine different patterns of lost information.

In the Hare example, we created the groups using the sum of all of the subscripts for the treatments. By doing so, we lose information on the five factor interaction effect ABCDE. We are using all of the subscripts so this effect involves all of the factors. The effect ABCDE also determines what other information we lose, i.e., it determines the pattern of aliasing. That will be discussed later in the section on aliasing.

EXAMPLE 10.2.2. *Groups for in a* 2^4 *experiment*
Consider now a 2^4 experiment with factors A, B, C, D each at two levels. We can use the *ABCD* effect to break the treatments into two groups.

ABCD Fractions

ABCD even	*ABCD* odd
$a_0b_0c_0d_0$	$a_0b_0c_0d_1$
$a_0b_0c_1d_1$	$a_0b_0c_1d_0$
$a_0b_1c_0d_1$	$a_0b_1c_0d_0$
$a_0b_1c_1d_0$	$a_0b_1c_1d_1$
$a_1b_0c_0d_1$	$a_1b_0c_0d_0$
$a_1b_0c_1d_0$	$a_1b_0c_1d_1$
$a_1b_1c_0d_0$	$a_1b_1c_0d_1$
$a_1b_1c_1d_1$	$a_1b_1c_1d_0$

On the left the sum of the subscripts for all of the treatments are even, e.g., the sum for treatment $a_0b_1c_1d_0$ is $0+1+1+0=2$ which is even. On the right, the sum of the subscripts is always odd.

We can use effects other than ABCD to break the treatments into two groups. If we only wanted two groups, it is unlikely that we would use any other effect, but to break things down into more than two groups it is important to understand this concept. For example, we can use the ABC effect to break the treatments into two groups. In this case, the groups will be determined by the sum of the subscripts on only the A, B, and C factors. The groups are

ABC Fractions

ABC even	ABC odd
$a_0b_0c_0d_0$	$a_0b_0c_1d_1$
$a_0b_1c_1d_0$	$a_0b_1c_0d_1$
$a_1b_0c_1d_1$	$a_1b_0c_0d_0$
$a_1b_1c_0d_1$	$a_1b_1c_1d_0$
$a_0b_0c_0d_1$	$a_0b_0c_1d_0$
$a_0b_1c_1d_1$	$a_0b_1c_0d_0$
$a_1b_0c_1d_0$	$a_1b_0c_0d_1$
$a_1b_1c_0d_0$	$a_1b_1c_1d_1$

On the left we have treatments like $a_0b_1c_1d_0$ with $0+1+1=2$ and on the right we have $a_0b_1c_0d_1$ with $0+1+0=1$. In both cases we have ignored the subscript on d when calculating the sum. Note that since we are ignoring the d subscript, if we have $a_0b_1c_1d_0$ on the left hand side, we also better have $a_0b_1c_1d_1$ on the left.

A final example uses BCD to break up the treatments. The groups are

BCD Fractions

BCD even	*BCD* odd
$a_0b_0c_0d_0$	$a_0b_0c_0d_1$
$a_0b_1c_1d_0$	$a_0b_1c_1d_1$
$a_1b_0c_1d_1$	$a_1b_0c_1d_0$
$a_1b_1c_0d_1$	$a_1b_1c_0d_0$
$a_0b_0c_1d_1$	$a_0b_0c_1d_0$
$a_0b_1c_0d_1$	$a_0b_1c_0d_0$
$a_1b_0c_0d_0$	$a_1b_0c_0d_1$
$a_1b_1c_1d_0$	$a_1b_1c_1d_1$

If we want to break the treatments up into four groups, we need to use two defining effects. For example, we can cross the groupings for ABC and the groupings for BCD. The table below gives four groups. Any one of the groups can be used as a $1/4$ replicate of a 2^4 experiment.

ABC, *BCD* Fractions

ABC even		*ABC* odd	
BCD even	*BCD* odd	*BCD* even	*BCD* odd
$a_0b_0c_0d_0$	$a_0b_0c_0d_1$	$a_0b_0c_1d_1$	$a_0b_0c_1d_0$
$a_0b_1c_1d_0$	$a_0b_1c_1d_1$	$a_0b_1c_0d_1$	$a_0b_1c_0d_0$
$a_1b_0c_1d_1$	$a_1b_0c_1d_0$	$a_1b_0c_0d_0$	$a_1b_0c_0d_1$
$a_1b_1c_0d_1$	$a_1b_1c_0d_0$	$a_1b_1c_1d_0$	$a_1b_1c_1d_1$

In the first column, the a, b, c subscripts add to an even number and the b, c, d subscripts also add to an even number. In the second column, the a, b, c subscripts add to an even number but the b, c, d subscripts add to an odd number. In the third column, the a, b, c subscripts sum to an odd, and the b, c, d subscripts sum to an even. In the last column, the a, b, c subscripts sum to odd and the b, c, d subscripts also sum to odd. □

10.3 Aliasing

As mentioned in the previous section, when we only look at some of the treatment combinations, we loose the ability to tell various treatment effects apart. We can tell which treatments are aliased with each other by performing a type of multiplication.

Consider a $1/2$ replicate of a 2^4 in which the groups are determined by the *ABCD* effect. To find aliases for an effect, we multiply the effect by *ABCD* and drop out any factors that get raised to an even power. Any factors that get raised to an odd power are just left in the product. For example, the effect A is aliased with $A \times ABCD = A^2BCD = BCD$. The effect BC is aliased with $BC \times ABCD = AB^2C^2D = AD$. The effect BCD is aliased with $BCD \times ABCD = AB^2C^2D^2 = A$. Oops, we already knew that!

The entire aliasing structure is given below.

Effect	$\times ABCD$	Alias
A	=	BCD
B	=	ACD
C	=	ABD
D	=	ABC
AB	=	CD
AC	=	BD
BC	=	AD
AD	=	BC
BD	=	AC
CD	=	AB
ABC	=	D
ABD	=	C
ACD	=	B
BCD	=	A
$ABCD$		—

Table 10.3: *Aliases for a* $1/2$ *rep. from a* 2^5 *based on ABCDE*

Effect	$\times ABCDE$	Alias
A	=	$BCDE$
B	=	$ACDE$
C	=	$ABDE$
D	=	$ABCE$
E	=	$ABCD$
AB	=	CDE
AC	=	BDE
AD	=	BCE
AE	=	BCD
BC	=	ADE
BD	=	ACE
BE	=	ACD
CD	=	ABE
CE	=	ABD
DE	=	ABC

If we consider the $1/2$ replicate of a 2^4 with groups determined by *BCD*, the aliasing structure is given below.

Effect	$\times BCD$	Alias
A	=	$ABCD$
B	=	CD
C	=	BD
D	=	BC
AB	=	ACD
AC	=	ABD
BC	=	AD
AD	=	ABC
BD	=	C
CD	=	B
ABC	=	AD
ABD	=	AC
ACD	=	AB
BCD		—
$ABCD$	=	A

For example $ABC \times BCD = AB^2C^2D = AD$.

Table 10.3 gives the aliases that apply to Hare's experiment.

When we create more than two groups in order to get something smaller than a $1/2$ replicate, the aliasing structure becomes more involved. Consider again the groups formed by *ABC* and *BCD* in a 2^4. The aliases involve both of these defining effects as well as another effect. For a $1/4$th replicate, we need 4 groups so there is going to be $4-1=3$ effects involved in aliasing. If we do a $1/8$ replication, this involves creating 8 groups, so there would be $8-1=7$ effects involved in aliasing. We need to be able to identify these other effects. With *ABC* and *BCD* defining groups, the other effect is $ABC \times BCD = AB^2C^2D = AD$. To see that this makes some sense, observe that we would get the same four groups if we used, say, ABC and AD to define the groups rather than ABC and BCD. As illustrated earlier, with ABC and BCD defining groups, the groups of treatments are as given below.

ABC even		*ABC* odd	
BCD even	*BCD* odd	*BCD* even	*BCD* odd
$a_0b_0c_0d_0$	$a_0b_0c_0d_1$	$a_0b_0c_1d_1$	$a_0b_0c_1d_0$
$a_0b_1c_1d_0$	$a_0b_1c_1d_1$	$a_0b_1c_0d_1$	$a_0b_1c_0d_0$
$a_1b_0c_1d_1$	$a_1b_0c_1d_0$	$a_1b_0c_0d_0$	$a_1b_0c_0d_1$
$a_1b_1c_0d_1$	$a_1b_1c_0d_0$	$a_1b_1c_1d_0$	$a_1b_1c_1d_1$

With ABC and AD defining the groups we get the same four groups.

ABC even		*ABC* odd	
AD even	*AD* odd	*AD* odd	*AD* even
$a_0b_0c_0d_0$	$a_0b_0c_0d_1$	$a_0b_0c_1d_1$	$a_0b_0c_1d_0$
$a_0b_1c_1d_0$	$a_0b_1c_1d_1$	$a_0b_1c_0d_1$	$a_0b_1c_0d_0$
$a_1b_0c_1d_1$	$a_1b_0c_1d_0$	$a_1b_0c_0d_0$	$a_1b_0c_0d_1$
$a_1b_1c_0d_1$	$a_1b_1c_0d_0$	$a_1b_1c_1d_0$	$a_1b_1c_1d_1$

We would also get the same four groups if we used AD and BCD to define the groups.

In terms of finding the aliases, the $1/4$ replication of a 2^4 involves collecting data on only four treatments, so there are only $4-1=3$ effects that can really be estimated. To find these, multiply each effect by all three of the defining effects, *ABC*, *BCD*, and *AD*.

$$\begin{aligned} A &\times ABC = A^2BC = BC, \\ A &\times BCD = ABCD, \\ A &\times AD = A^2D = D, \end{aligned}$$

so

$$A = BC = ABCD = D.$$

Similarly,

$$B = AC = CD = ABD$$

and

$$C = AB = BD = ACD.$$

EXAMPLE 10.3.3. $1/8$ *replication of a* 2^8 *experiment*
The 2^8 experiment involves eight factors, call them A through H. A $1/8$th replication of the $2^8 = 256$ treatments involves only 32 treatments. The $1/8 = 2^{-3}$ replication involves specifying 3 defining effects, say *ABCD*, *EFGH*, and *CDEF* but with 8 groups there are really $8-1=7$ effects involved in aliasing. Multiplying pairs of these defining effects and also multiplying all three of the effects together give the 4 other effects that implicitly define the $1/8$ replication. These other implicit defining effects are *ABEF*, *CDGH*, *ABCDEFGH*, and *ABGH*. For example,

$$ABCD \times EFGH = ABCDEFGH.$$

To find the alias of an effect, multiply the effect by all 7 of these aliasing effects. For A the aliases are

$$\begin{aligned} A = A(ABCD) = A(EFGH) = A(CDEF) = A(ABED) \\ = A(CDGH) = A(ABCDEFGH) = A(ABGH). \end{aligned}$$

Simplifying gives

$$A = BCD = AEFGH = ACDEF = BED = ACDGH = BCDEFGH = BGH.$$

The two-factor effect *AB* has aliases

$$\begin{aligned} AB = AB(ABCD) = AB(EFGH) = AB(CDEF) = AB(ABED) \\ = AB(CDGH) = AB(ABCDEFGH) = AB(ABGH) \end{aligned}$$

which upon simplification become

$$AB = CD = ABEFGH = ABCDEF = ED = ABCDGH = CDEFGH = GH.$$

Table 10.4: *Hare's* $1/2$ *rep. from a* 2^5 *based on ABCDE*

Batch	Treatment	s_c	s_p
1	$a_0b_0c_0d_1e_1$	0.43	0.78
2	$a_1b_0c_1d_1e_1$	0.52	1.10
3	$a_1b_1c_0d_0e_0$	0.58	1.70
4	$a_1b_0c_1d_0e_0$	0.55	1.28
5	$a_0b_1c_0d_0e_1$	0.58	0.97
6	$a_0b_0c_1d_0e_1$	0.60	1.47
7	$a_0b_1c_0d_1e_0$	1.04	1.85
8	$a_1b_1c_1d_1e_0$	0.53	2.10
9	$a_0b_1c_1d_1e_1$	0.38	0.76
10	$a_1b_1c_0d_1e_1$	0.41	0.62
11	$a_0b_0c_1d_1e_0$	0.66	1.09
12	$a_0b_0c_0d_0e_0$	0.55	1.13
13	$a_1b_0c_0d_0e_1$	0.65	1.25
14	$a_1b_1c_1d_0e_1$	0.72	0.98
15	$a_1b_0c_0d_1e_0$	0.48	1.36
16	$a_0b_1c_1d_0e_0$	0.68	1.18

10.4 Analysis Methods

One new problem we have when fitting ANOVA or other models to fractional replications is that there is no natural estimate of error because there is no replication. We don't even have observations on every factor combination, much less multiple observations on treatments. We present two ways to proceed. One is to assume that higher-order interactions do not exist and use them to estimate the error. The other is based on a graphical display of the effects that is similar in spirit to a normal plot.

EXAMPLE 10.4.1. Consider again the $1/2$ rep. of a 2^5 from Hare (1988). The experimental background was discussed in Example 10.1.5. The two $1/2$ rep. treatment groups were given in Table 10.2 and the aliasing structure was given in Table 10.3. The issue is excessive variability in the taste of a dry soup mix due to the 'intermix' containing flavorful ingredients such as salt and vegetable oil. For each intermix batch (treatment combination), the original data are groups of 5 samples taken every 15 minutes throughout a day of processing. Thus each batch yields data for a balanced one-way analysis of variance with $N = 5$. The data actually analyzed are derived from the ANOVAs on different batches. There are two sources of variability in the original observations, the variability within a group of 5 samples and variability that occurs between 15 minute intervals. From the analysis of variance data, the within group variability is estimated with the *MSE* and summarized as the estimated 'capability' standard deviation

$$s_c = \sqrt{MSE}.$$

The 'process' standard deviation was defined as the standard deviation of an individual observation. The standard deviation of an observation incorporates both the between group and the within group sources of variability. Based on a one-way ANOVA model in which the group means are considered random, the estimated process standard deviation is taken as

$$s_p = \sqrt{MSE + \frac{MSGrps - MSE}{5}},$$

where the 5 is the number of samples taken at each time. These two statistics, s_c and s_p, are available from every batch of soup mix prepared and provide the data for analyzing batches. The $1/2$ rep. of a 2^5 specifies different ways of making batches of soup mix. The design and the standard deviations are given in Table 10.4. For now, we analyze only the data on s_p.

This is a "resolution V" (V is a Roman numeral) design because the only defining effect for

Table 10.5: *ANOVA for Hare's s_p*

Source	Source	*df*	*SS*	Rank
A	*A*	1	0.0841	10
B	*B*	1	0.0306	7
C	*C*	1	0.0056	4
D	*D*	1	0.0056	3
AB	*AB*	1	0.0009	1
AC	*AC*	1	0.0361	8
AD	*AD*	1	0.0036	2
BC	*BC*	1	0.0182	5
BD	*BD*	1	0.1056	12
CD	*CD*	1	0.0210	6
ABC	*DE*	1	0.3969	13
ABD	*CE*	1	0.0729	9
ACD	*BE*	1	0.6561	14
BCD	*AE*	1	0.0930	11
ABCD	*E*	1	0.8836	15
Total	Total	15	2.4140	

the fractional replication is a 5-factor interaction. Given our methods for finding aliases, all main effects are confounded with four-factor interactions and all two-factor interactions are confounded with three-factor interactions. If we are prepared to assume that there are no three- or four-factor interactions, we have available estimates of all the main effects and two-factor interactions. Table 10.5 contains an ANOVA table with two columns labeled 'Source.' The first does not involve factor *E* and the second replaces high-order interactions not involving E with their lower order aliases. Table 10.5 also contains a ranking of the sizes of the sums of squares from smallest to largest.

The reason for having two Source columns in Table 10.5 is that many ANOVA computer programs will not fit fractional factorials. If you are stuck with such a program, the simplest way to obtain an analysis is to trick it into doing most of the work. If we could drop one of our factors, our 1/2 rep. would become a full factorial (without replication) on the remaining factors. For example, *if we dropped factor E, and thus dropped the e terms from all the treatment combinations in Table 10.4, we would have observations on all 16 of the treatment combinations in the* 2^4 *defined by A, B, C, and D*. It is easy to find computer programs that will analyze a full factorial. Table 10.5 gives the results of an analysis in which we ignored the presence of factor *E*. The first column of Table 10.5 contains the sources from the full factorial on *A*, *B*, *C*, and *D*; the second column replaces the higher order interactions from the full factorial with their lower order aliases. In computing the full factorial one might also have to drop the four-factor interaction from the model so that the program can relabel it as a one degree of freedom Error term.

The simpler, but more dangerous, method of analysis is to assume that no higher order interactions exist and form an error term by pooling the estimable terms that involve only higher order interactions. A particular term involves only higher order interactions if the term and all of its aliases are high order interactions. What we mean by high order interactions is intentionally left ill defined to maintain flexibility. In this design, unless you consider second-order interactions as higher order, there are no terms involving only higher order interactions. Most often, higher order interactions are taken to be interactions that only involve three or more factors, but in designs like this, one *might* be willing to consider two-factor interactions as higher order to obtain an error term for testing main effects. (I personally would not be willing to do it with these data.) Often terms that involve only three and higher order interactions are pooled into an error, but in designs with more factors and many high order interactions, one might wish to estimate three-factor interactions and use only terms involving four or more factors in a pooled error.

If we assume away all two-factor and higher order interactions for the present data, the ANOVA table becomes that displayed in Table 10.6. With this error term, only factor *E* appears to be important. As we will see later, most of the important effects in these data seem to be interactions, so the

Table 10.6: *Analysis of variance on s_p for Hare's data*

Source	df	SS	MS	F
A	1	0.0841	0.0841	0.60
B	1	0.0306	0.0306	0.22
C	1	0.0056	0.0056	0.04
D	1	0.0056	0.0056	0.04
E	1	0.8836	0.8836	6.29
Error	10	1.4044	0.1404	
Total	15	2.4140		

error term based on no interactions is probably inappropriate. That is why we referred to this as the simpler but more dangerous method. □

Graphical methods of analysis are similar to normal plots and are based on plotting ordered (ranked) sums of squares. In a normal plot, the data from a single sample are ordered from smallest to largest and plotted against the *expected order statistics* from a standard normal distribution. In other words, the smallest observation in a sample of size, say, 13 is plotted against the expected value for the smallest observation in a sample of size 13 from a $N(0,1)$ distribution. The second smallest observation is plotted against the expected value for the second smallest observation in a sample of size 13 from a $N(0,1)$ distribution, and so on. This plot should approximate a straight line if the data are truly normal, the slope of the plot estimates the standard deviation of the population, and the intercept estimates the population mean.

The graphical display presented here is a $\chi^2(1)$ plot. In general, find the sums of squares for the various factors in the 2^{f-r} experiment and denote them SS_j. If there are no effects in the experiment, every SS_j should have a $\chi^2(1)$ distribution multiplied by σ^2. Any term that corresponds to a substantial effect should be larger than a $\sigma^2\chi^2(1)$ distribution. Order the sums of squares from smallest to largest denoted $SS_{(j)}$. We are going to plot these order statistics against the expected values of the order statistics from a $\chi^2(1)$ distribution. If there are no effects in the experiment, this should give a rough straight line (with slope σ^2). Large effects should be visible in that they should correspond to the largest order statistics and should jump up from the straight line created by the plot of the smaller expected order statistics against the smaller observed order statistics. It is difficult to find the expected order statistics but they are easy to approximate. If F is the the cdf for the $\chi^2(1)$ distribution, the expected value of the jth order statistic is approximately $F^{-1}[j/(n+1)]$. Thus we plot the pairs $\left(F^{-1}[j/(n+1)], SS_{(j)}\right)$, looking for a linear relationship for the smaller values and with upward jumps from the linear relationship identifying important effects.

A more common method than to use $\chi^2(1)$ plots is to use half-normal plots of the positive square roots of the sums of squares. To construct half-normal plots requires the cdf for the half-normal distribution which is $p = 2\Phi(x) - 1$ for $x > 0$. Here Φ is the cdf of a standard normal distribution. The inverse of the half-normal cdf is $x = \Phi^{-1}\left(\frac{p+1}{2}\right)$. The half normal plot is of the pairs $\left(\Phi^{-1}\left[\frac{p_j+1}{2}\right], \sqrt{SS_{(j)}}\right)$ where, by analogy with the $\chi^2(1)$ plot, we would use $p_j = j/(n+1)$, but more commonly one takes $p_j = (j-0.5)/n$. Again, this should be a rough straight line with important effects jumping up from the line on the right side.

EXAMPLE 10.4.1 CONTINUED. From the ranked sums of squares in Table 10.5 and the formula for computing approximate $\chi^2(1)$ scores, we construct Table 10.7 containing the scores and the ordered sums of squares necessary for the $\chi^2(1)$ plot of the 15 effects from Hare's data. Figure 10.1 contains the plot. Again, the $\chi^2(1)$ scores in Table 10.13 are approximate expected order statistics. They are computed by applying the inverse of the $\chi^2(1)$ cumulative distribution function to the values $j/(n+1)$, where j goes from 1 to 15 and $n = 15$.

The key to the graphical analysis is that nonnegligible treatment effects cause the sums of

Table 10.7: $\chi^2(1)$ *scores and ordered sums of squares for Hare's (1988) data.*

$\chi^2(1)$ scores	Ordered *SS*
0.00615	0.0009
0.02475	0.0036
0.07711	0.0056
0.07711	0.0056
0.16181	0.0182
0.23890	0.0210
0.33539	0.0306
0.45494	0.0361
0.60283	0.0729
0.78703	0.0841
1.02008	0.0930
1.32330	0.1056
1.73715	0.3969
2.35353	0.6561
3.46977	0.8836

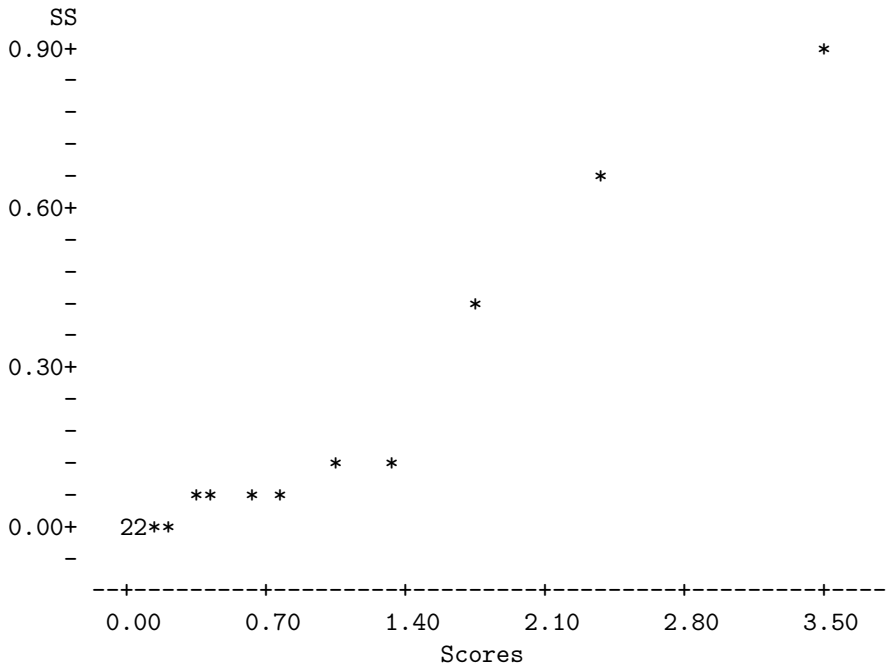

Figure 10.1: $\chi^2(1)$ *plot of sums of squares.*

squares to estimate something larger than σ^2. The sums of squares for nonnegligible effects should show up in the plot as inappropriately large values. The lower 12 observations in Figure 10.1 seem to fit roughly on a line, but the three largest observations seem to be inconsistent with the others. These three observations correspond to the most important effects in the data. From the rankings in Table 10.5, we see that the important effects are E, BE, and DE.

We need to evaluate the meaning of the important effects. The largest effect is due to E, the delay in using the intermix. However, this effect is complicated by interactions involving the delay. To evaluate the BE interaction we need the means for the four combinations of B and E.

	B	
E	b_0	b_1
e_0	1.215	1.7075
e_1	1.150	0.8325

The BE interaction is due to the fact that running the mixer at room temperature, b_1, increases variability if the intermix is used after one day, e_0, but decreases variability if the intermix is used a week later, e_1. However, the variability under delay is smaller for both B levels than the variability for immediate use with either B level. This suggests delaying use of the intermix.

To evaluate the DE interaction we need the means for the four combinations of D and E.

	D	
E	d_0	d_1
e_0	1.3225	1.6000
e_1	1.1675	0.8150

A large batch weight, d_1, causes increased variability when the intermix is used immediately but decreased variability when use is delayed to 7 days. Again, it is uniformly better to delay.

Overall, the best combination of treatments involve e_1, b_1 and d_1. Recalling that the current process is batch 7 with one vegetable oil port, room temperature mixing, 60 seconds mixing time, 2000 pound batches, and a 1 day delay, we would recommend changing to a 7 day delay because that minimizes process variability. As alluded to in the next chapter, if this change for the purpose of reducing process variability puts the process off target, it might be possible to use factors A and C to put it back on target, since A and C do not seem to affect process variability. □

Christensen (1996, Sections 17.3 and 17.4) expands on this graphical analysis.

Minitab commands

Below are given Minitab commands for obtaining the analysis given. The data file had eight columns, the first six were indicators for batch and factors A, B, C, D and E, respectively. Columns 7 and 8 contained the data on s_c and s_p.

```
names c8 'y' c2 'a' c3 'b' c4 'c' c5 'd' c6 'e'
anova c8=c2|c3|c4|c5 - c2*c3*c4*c5
note     AFTER SEEING THE ANOVA, ENTER THE SUMS
note     OF SQUARES INTO c10.
set c10
841 306 56 56 9 361 36 182
1056 210 3969 729 6561 930 8836
end
let c10=c10/10000
note     CONSTRUCT CHI-SQUARED SCORES AND PLOT.
rank c10 c11
let c11=c11/16
invcdf c11 c12;
chisquare 1.
plot c10 c12
```

Note that $c6$ was not used in the anova command. Factor E was dropped to deal with the fractional nature of the factorial. Minitab's ANOVA command requires an error term to exist in the model. The command given above specifies a full factorial model ($c2|c3|c4|c5$) but subtracts out the $ABCD$

Table 10.8: *Alternative forms of identifying treatments: subscripts only and only letters with subscript 1.*

Treatment Notation			
Treatment	Subscripts only	Letters with $_1$.	Yield
$a_0b_0c_0d_0e_0f_0g_0h_0j_0$	000000000	$---$	17
$a_1b_0c_0d_0e_0f_1g_1h_1j_1$	100001111	$afghj$	21
$a_0b_1c_0d_0e_1f_0g_1h_1j_1$	010010111	$beghj$	41
$a_1b_1c_0d_0e_1f_1g_0h_0j_0$	110011000	$abef$	20
$a_0b_0c_1d_0e_1f_1g_0h_0j_1$	001011001	$cefj$	10
$a_1b_0c_1d_0e_1f_0g_1h_1j_0$	101010110	$acegh$	42
$a_0b_1c_1d_0e_0f_1g_1h_1j_0$	011001110	$bcfgh$	14
$a_1b_1c_1d_0e_0f_0g_0h_0j_1$	111000001	$abcj$	58
$a_0b_0c_0d_1e_1f_1g_0h_1j_0$	000111010	$defh$	8
$a_1b_0c_0d_1e_1f_0g_1h_0j_1$	100110101	$adegj$	18
$a_0b_1c_0d_1e_0f_1g_1h_0j_1$	010101101	$bdfgj$	7
$a_1b_1c_0d_1e_0f_0g_0h_1j_0$	110100010	$abdh$	15
$a_0b_0c_1d_1e_0f_0g_0h_1j_1$	001100011	$cdhj$	8
$a_1b_0c_1d_1e_0f_1g_1h_0j_0$	101101100	$acdfg$	10
$a_0b_1c_1d_1e_1f_0g_1h_0j_0$	011110100	$bcdeg$	12
$a_1b_1c_1d_1e_1f_1g_0h_1j_1$	111111011	$abcdefghj$	10

interaction $(c2*c3*c4*c5)$ and sets it equal to the error. Thus, Minitab's error term is actually the *ABCD* interaction. The command 'set $c10$' is used to create a data column that contains the sums of squares for the various effects. The commands involving $c11$ and $c12$ are used to get the approximate expected order statistics from a $\chi^2(1)$ and to plot the ordered sums of squares against the expected order statistics. After identifying the important effects, the ANOVA command can be repeated with various factors deleted to obtain the necessary means tables.

10.5 Alternative forms of identifying treatments

In this chapter we have used the subscripts 0 and 1 to indicate the levels of treatments. Alternative notations are extremely common and must be recognized. Tables 10.4 through 10.6 provide some alternative notations for the 16 treatments actually used in the viscosity experiment of Example 10.1.1. Table 10.4 provides two alternatives and the actual data. The first alternative relies on the fact that if you know the order of the factors, all you really need are the subscript values. The second alternative only reports the letters with the subscript 1. Table 10.5 replaces the subscript 0 with -1 and reports only the subscripts. Here we have explicitly written 1 as $+1$, but the $+$ sign is irrelevant! The notation using -1 and 1 is quite common, is particularly useful for response surface designs, and is employed in Table 11.4. Table 9.21 also used the -1 and 1 notation and Tables 9.16 and 9.19 used -1, 0, 1 to denote three levels of a factor. Table 10.6 replaces -1 and $+1$ with $-$ and $+$.

10.6 Placket-Burman Designs

When looking at 2^f designs, a generalization of looking at 2^{f-r} fractional factorials is the use of Placket-Burman designs. They are based on Hadamard matrices and exist for some run sizes that are multiples of $n = 4$. These include $n = 4, 8, 12, 16, 20$. Designs with $n = 12, 20$ are not powers of 2, so they add flexibility to the choice of designs beyond the 2^{f-r} fractional designs. For example you can estimate main effects for 9, 10, or 11 factors in 12 runs with a Placket-Burman design whereas the smallest 2^{f-r} design involves $16 = 2^{9-5} = 2^{10-6} = 2^{11-7}$ different runs. *TiD* contains more information on Placket-Burman designs.

Table 10.9: *Alternative forms of identifying treatments: subscripts only, 1,0 replaced with* ± 1.

Treatment	A	B	C	D	E	F	G	H	J
$a_0b_0c_0d_0e_0f_0g_0h_0j_0$	-1	-1	-1	-1	-1	-1	-1	-1	-1
$a_1b_0c_0d_0e_0f_1g_1h_1j_1$	$+1$	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$
$a_0b_1c_0d_0e_1f_0g_1h_1j_1$	-1	$+1$	-1	-1	$+1$	-1	$+1$	$+1$	$+1$
$a_1b_1c_0d_0e_1f_1g_0h_0j_0$	$+1$	$+1$	-1	-1	$+1$	$+1$	-1	-1	-1
$a_0b_0c_1d_0e_1f_1g_0h_0j_1$	-1	-1	$+1$	-1	$+1$	$+1$	-1	-1	$+1$
$a_1b_0c_1d_0e_1f_0g_1h_1j_0$	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	$+1$	-1
$a_0b_1c_1d_0e_0f_1g_1h_1j_0$	-1	$+1$	$+1$	-1	-1	$+1$	$+1$	$+1$	-1
$a_1b_1c_1d_0e_0f_0g_0h_0j_1$	$+1$	$+1$	$+1$	-1	-1	-1	-1	-1	$+1$
$a_0b_0c_0d_1e_1f_1g_0h_1j_0$	-1	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	-1
$a_1b_0c_0d_1e_1f_0g_1h_0j_1$	$+1$	-1	-1	$+1$	$+1$	-1	$+1$	-1	$+1$
$a_0b_1c_0d_1e_0f_1g_1h_0j_1$	-1	$+1$	-1	$+1$	-1	$+1$	$+1$	-1	$+1$
$a_1b_1c_0d_1e_0f_0g_0h_1j_0$	$+1$	$+1$	-1	$+1$	-1	-1	-1	$+1$	-1
$a_0b_0c_1d_1e_0f_0g_0h_1j_1$	-1	-1	$+1$	$+1$	-1	-1	-1	$+1$	$+1$
$a_1b_0c_1d_1e_0f_1g_1h_0j_0$	$+1$	-1	$+1$	$+1$	-1	$+1$	$+1$	-1	-1
$a_0b_1c_1d_1e_0f_0g_1h_0j_0$	-1	$+1$	$+1$	$+1$	$+1$	-1	$+1$	-1	-1
$a_1b_1c_1d_1e_1f_1g_0h_1j_1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	-1	$+1$	$+1$

Table 10.10: *Alternative forms of identifying treatments: subscripts only, 0,1 replaced with* $-$, $+$.

Treatment	A	B	C	D	E	F	G	H	J
$a_0b_0c_0d_0e_0f_0g_0h_0j_0$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
$a_1b_0c_0d_0e_0f_1g_1h_1j_1$	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$
$a_0b_1c_0d_0e_1f_0g_1h_1j_1$	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$+$	$+$
$a_1b_1c_0d_0e_1f_1g_0h_0j_0$	$+$	$+$	$-$	$-$	$+$	$+$	$-$	$-$	$-$
$a_0b_0c_1d_0e_1f_1g_0h_0j_1$	$-$	$-$	$+$	$-$	$+$	$+$	$-$	$-$	$+$
$a_1b_0c_1d_0e_1f_0g_1h_1j_0$	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$+$	$-$
$a_0b_1c_1d_0e_0f_1g_1h_1j_0$	$-$	$+$	$+$	$-$	$-$	$+$	$+$	$+$	$-$
$a_1b_1c_1d_0e_0f_0g_0h_0j_1$	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$-$	$+$
$a_0b_0c_0d_1e_1f_1g_0h_1j_0$	$-$	$-$	$-$	$+$	$+$	$+$	$-$	$+$	$-$
$a_1b_0c_0d_1e_1f_0g_1h_0j_1$	$+$	$-$	$-$	$+$	$+$	$-$	$+$	$-$	$+$
$a_0b_1c_0d_1e_0f_1g_0h_1j_1$	$-$	$+$	$-$	$+$	$-$	$+$	$+$	$-$	$+$
$a_1b_1c_0d_1e_0f_0g_0h_1j_0$	$+$	$+$	$-$	$+$	$-$	$-$	$-$	$+$	$-$
$a_0b_0c_1d_1e_0f_0g_0h_1j_1$	$-$	$-$	$+$	$+$	$-$	$-$	$-$	$+$	$+$
$a_1b_0c_1d_1e_0f_1g_1h_0j_0$	$+$	$-$	$+$	$+$	$-$	$+$	$+$	$-$	$-$
$a_0b_1c_1d_1e_1f_0g_1h_0j_0$	$-$	$+$	$+$	$+$	$+$	$-$	$+$	$-$	$-$
$a_1b_1c_1d_1e_1f_1g_0h_1j_1$	$+$	$+$	$+$	$+$	$+$	$+$	$-$	$+$	$+$

10.7 Exercises

EXERCISE 10.7.1. In Example 10.1.1. we introduced the viscosity example of Taguchi (1987) and Box and Bisgaard (1994). The 16 treatments were obtained by using the defining effects

$$BCDE, \quad ABG, \quad ABCJ, \quad ABDH, \quad ABEF.$$

(These are hard to guess but easy to check.) Note that with 9 factors and 5 defining effects, the number of treatments being examined is $16 = 2^4 = 2^9/2^5 = 2^{9-5}$. The aliasing structure here is very complicated because in addition to the 5 defining effects, there are 26 other effects involved in the aliasing. There are $5 = \binom{5}{1}$ original defining effects, $10 = \binom{5}{2}$ effects obtained by multiplying pairs of the defining effects, $10 = \binom{5}{3}$ obtained by multiplying triples of the defining effects, $5 = \binom{5}{4}$ obtained by multiplying groups of 4 defining effects, and $1 = \binom{5}{5}$ effect obtained by multiplying together all five defining effects. The actual treatments and data are given in Table 10.4. After evaluating a residual analysis on a reduced model for the untransformed data, Box and Bisgaard (1994) suggest taking logs of the data.

Table 10.11: *Wave soldering.*

	treatment subscripts								
Case	a	b	c	d	e	f	g	h	Yellow
1	0	0	0	0	0	0	0	0	6.00
2	1	0	0	0	0	0	1	1	10.00
3	0	1	0	0	0	0	0	1	10.00
4	1	1	0	0	0	0	1	0	8.50
5	0	0	1	0	0	1	0	1	1.50
6	1	0	1	0	0	1	1	0	0.25
7	0	1	1	0	0	1	0	0	1.75
8	1	1	1	0	0	1	1	1	4.25
9	0	0	0	1	0	1	1	1	6.50
10	1	0	0	1	0	1	0	0	0.75
11	0	1	0	1	0	1	1	0	3.50
12	1	1	0	1	0	1	0	1	3.25
13	0	0	1	1	0	0	1	0	6.00
14	1	0	1	1	0	0	0	1	9.50
15	0	1	1	1	0	0	1	1	6.25
16	1	1	1	1	0	0	0	0	6.75
17	0	0	0	0	1	0	0	0	20.00
18	1	0	0	0	1	0	1	1	16.50
19	0	1	0	0	1	0	0	1	17.25
20	1	1	0	0	1	0	1	0	19.50
21	0	0	1	0	1	1	0	1	9.67
22	1	0	1	0	1	1	1	0	2.00
23	0	1	1	0	1	1	0	0	5.67
24	1	1	1	0	1	1	1	1	3.75
25	0	0	0	1	1	1	1	1	6.00
26	1	0	0	1	1	1	0	0	7.30
27	0	1	0	1	1	1	1	0	8.70
28	1	1	0	1	1	1	0	1	9.00
29	0	0	1	1	1	0	1	0	19.30
30	1	0	1	1	1	0	0	1	26.70
31	0	1	1	1	1	0	1	1	17.70
32	1	1	1	1	1	0	0	0	10.30

EXERCISE 10.7.2. In Example 10.1.2. we introduced the Shina (1991) and Lawson and Helps (1996) [q96-467.dat] experiment on wave-soldering with 8 factors but only $32 = 2^{8-3}$ observations. The effects involved in aliasing are

$$ABFH, \quad ACFG, \quad ADG, \quad BCGH, \quad BDFGH, \quad CDF, \quad ABCDH.$$

The first three are defining effects. The actual design and data are given in Table 10.7. Yellow is the dependent variable.

EXERCISE 10.7.3. In Example 10.1.3. we introduced Sarkar's (1997) [qe97-529.dat] 2^{5-1} experiment on cleaning efficiencies with five factors but only 16 observations. Tables 10.8 and 10.9 give the design and the data, respectively.

Do you find the number of identical data values disturbing? Are there two dependent variables? Why are there 4 replications? Are they studying variability?

EXERCISE 10.7.4. In Example 10.1.4 we considered an unreplicated 2^4 design given in Ferrer and Romero (1995) [qe95-750.dat] for the improvement of adhesive force between polyurethane sheets. The design and data are given in Table 10.10. Use these data to illustrate the method of analysis for experiments that do not involve replications.

EXERCISE 10.7.5. Chapman (1996) data from qe96-35.dat in Table 10.11. Photographic color

Table 10.12: *Sarkar's design.*

Trial	ERM Feed Rate a	MCO Setting Angle b	MCO Setting Distance c	ERM Feed Setting d	ERM Opener Setting e
1	40	6	5	4.5	4
2	40	6	5	3	3
3	40	6	4	4.5	3
4	40	6	4	3	4
5	40	10	5	4.5	3
6	40	10	5	3	4
7	40	10	4	4.5	4
8	40	10	4	3	3
9	72	6	5	4.5	3
10	72	6	5	3	4
11	72	6	4	4.5	4
12	72	6	4	3	3
13	72	10	5	4.5	4
14	72	10	5	3	3
15	72	10	4	4.5	3
16	72	10	4	3	4

Table 10.13: *Sarkar's data.*

	Cleaning Efficiency							
	After ERM Opener				Overall			
Trial	1	2	3	4	1	2	3	4
1	35.99	38.33	38.72	39.89	51.20	51.20	52.38	51.99
2	38.33	39.11	39.89	38.72	50.81	50.42	51.20	50.81
3	39.89	39.42	39.11	38.72	51.20	50.81	50.81	50.42
4	38.72	38.72	39.11	39.42	51.20	50.81	50.81	50.42
5	39.89	39.89	39.11	39.42	51.20	50.81	50.81	50.42
6	39.42	39.42	39.11	38.72	51.20	50.20	50.81	50.42
7	40.28	39.11	39.11	38.72	51.20	50.81	50.81	50.42
8	39.11	39.11	39.42	39.21	51.20	51.20	50.81	50.81
9	55.06	55.06	50.30	52.30	56.35	58.08	57.64	56.78
10	55.06	54.19	53.76	52.46	57.22	57.22	56.78	56.78
11	55.92	55.49	55.49	53.76	57.64	57.64	57.22	56.78
12	54.19	53.76	53.76	53.33	57.22	56.35	56.68	55.92
13	55.92	55.06	54.62	54.19	58.08	57.64	57.22	56.35
14	55.06	54.62	54.62	54.19	57.64	57.64	57.22	57.78
15	54.62	54.62	54.19	54.19	57.22	57.78	57.78	55.92
16	54.62	54.19	54.19	53.76	56.78	56.35	56.35	55.92

slide development. Responses refer to levels of Red, Green, and Blue. Design variables are 6 developer constituents.

EXERCISE 10.7.7. Anand (1994) data from qe94-39.dat in Table 10.12.

Table 10.14: *Ferrer and Romero's adhesive force data.*

	treatment subscripts				Adhesive
Case	a	b	c	d	Force
1	0	0	0	0	3.80
2	1	0	0	0	4.34
3	0	1	0	0	3.54
4	1	1	0	0	4.59
5	0	0	1	0	3.95
6	1	0	1	0	4.83
7	0	1	1	0	4.86
8	1	1	1	0	5.28
9	0	0	0	1	3.29
10	1	0	0	1	2.82
11	0	1	0	1	4.59
12	1	1	0	1	4.68
13	0	0	1	1	2.73
14	1	0	1	1	4.31
15	0	1	1	1	5.16
16	1	1	1	1	6.06

Table 10.15: *Chapman's screening experiment for 12 variables*

D1	D2	D3	D4	D5	D6	Rmx	Gmx	Bmx	Rhd	Ghd	Bhd
1.5	2.1	7	.2	6.8	9.75	4	−8	−5	−1	0	−3
7.0	2.1	7	.1	3.5	9.75	−4	−17	−21	−6	−10	−13
1.5	3.6	7	.1	6.8	9.55	11	2	7	11	11	8
7.0	3.6	7	.2	3.5	9.55	2	−9	−2	7	3	1
1.5	2.1	18	.2	3.5	9.55	3	−5	−1	7	7	3
7.0	2.1	18	.1	6.8	9.55	0	−12	−4	4	4	4
1.5	3.6	18	.1	3.5	9.75	−2	−13	−3	−3	−4	−4
7.0	3.6	18	.2	6.8	9.75	−5	−22	−12	−4	−8	−8

D1	D2	D3	D4	D5	D6	Rld	Gld	Bld	Rmn	Gmn	Bmn
1.5	2.1	7	.2	6.8	9.75	5	7	5	5	4	7
7.0	2.1	7	.1	3.5	9.75	3	3	3	7	6	9
1.5	3.6	7	.1	6.8	9.55	13	16	14	7	6	9
7.0	3.6	7	.2	3.5	9.55	10	12	9	6	5	7
1.5	2.1	18	.2	3.5	9.55	11	16	13	7	6	8
7.0	2.1	18	.1	6.8	9.55	8	12	11	6	6	9
1.5	3.6	18	.1	3.5	9.75	3	5	4	6	4	7
7.0	3.6	18	.2	6.8	9.75	5	5	5	5	5	7

Table 10.16: *Anand's design and yields of deoiled wax*

	Temp					Slack	Deoiled			
Case	A	B	C	D	E	Wax	Wax	Yield	1	2
1	65	20	10	6	0	6.5	60.8	63.00	2.80	3.05
2	65	20	7	6	1	6.0	62.2	64.79	2.85	3.19
3	65	28	7	3	0	6.0	59.3	61.77	2.70	3.10
4	65	28	10	3	1	6.0	58.5	60.93	2.70	2.90
5	65	28	10	6	2	5.0	57.0	60.00	2.55	2.80
6	65	28	7	6	1	6.5	59.5	61.65	2.90	3.10
7	65	20	7	3	2	8.0	62.0	63.27	2.95	3.19
8	65	20	10	3	1	5.5	60.0	62.82	2.76	3.15
9	55	20	10	6	0	5.0	64.5	67.89	3.53	3.73
10	55	20	7	6	1	6.5	69.5	72.02	3.28	3.45
11	55	28	7	3	0	6.5	72.5	75.13	3.45	3.20
12	55	28	10	3	1	7.0	68.0	70.10	3.28	3.06
13	55	28	10	6	2	10.5	68.0	67.67	3.12	2.95
14	55	28	7	6	1	6.0	69.5	72.40	3.12	3.25
15	55	20	7	3	2	6.5	68.0	70.47	3.19	3.32
16	55	20	10	3	1	6.0	64.0	66.67	3.40	3.25

Chapter 11

Taguchi Methods

This chapter is still incomplete. Substantial parts are just copied from Christensen (2017, Section 3.4) and substantial parts of Christensen (2017, Section 3.4) were copied from here. Christensen (2017, Section 3.4) is far more complete, but at least some of the material there requires more technical background than is provided here.

As discussed in Chapter 1, high quality is less about achieving high performance than it is about maintaining consistency. One can usually improve performance with greater cost and associated diminishing returns for that cost. Quality improvement is more concerned with achieving the same (or better) performance with greater consistency and less cost. If someone doesn't like a product (the performance), they probably won't buy it in the first place. High quality is more about customer satisfaction, e.g., eliminating lemons.

For example, if you buy a car, you probably evaluate its rate of acceleration. When you pull out to pass another vehicle, with a semi coming the other way, you want that known rate of acceleration. You want it regardless of whether it is hot outside or cold; regardless of whether it is wet or dry. At the other end of the acceleration spectrum, if you are driving near a police officer, once again you probably don't want any surprises about the rate at which your vehicle accelerates.

Another example is the warm idle speed of an automobile. If it is too fast, the engine races, possibly damaging the engine, and making it more difficult to stop a car with an automatic transmission. On the other hand, if it idles too slowly, the engine dies. Again, you want an idle speed within certain tolerances regardless of the weather and regardless of the altitude at which you are driving. Uncontrollable factors that can affect performance are called *noise* factors. We are interested in reducing the ability of noise factors to affect performance.

How well a complicated machine performs will depend on how well individual parts conform to specifications. We want products that both hit the target specification on average and do so with as little variability as possible. This chapter is about methods of achieving that goal. In a famous example, Taguchi and Wu (1980) illustrated that increasing from 1% to 5% the lime content of clay used to make ceramic tiles, reduced the variability of tile sizes by a factor of 10.

Genichi Taguchi's major contribution to quality improvement was his emphasis on the importance of producing products that have small variability. His efforts in this regard would not have been successful without also providing usable methods for reducing variability. He placed emphasis on minimizing squared error losses from the target. He provided a very specific list of useful experimental designs labeled L_q where q is the number of treatments that you want to use. And he advocated the use of various signal to noise ratios in analyzing data. Christensen (2017), hereafter referred to as *TiD*, in Section 3.4 discusses his list of designs and relates them to more traditional developments in statistical experimental design. It also defines his signal to noise ratios and discusses their use, see Sections 2 and 3 below.

Taguchi's methods have been subjected to criticism. Myers, Khuri, and Vining (1992) summarize these criticisms as (a) inefficiency of the response measures used, (b) lack of flexibility in modeling, (c) lack of economy in experimental design, (d) overemphasis on optimization relative to understanding, and (e) overlooking the sequential nature of experimentation. Stephens (1995, 1996)

Table 11.1: *Taguchi L_9 ($3 \times 3 \times 3 \times 3$ Fractional Design). (qe94-X.dat)*

Run	a	b	c	d	Observations							
1	0	0	0	0	15.6	9.5	16.9	19.9	19.6	19.6	20.0	19.1
4	0	1	1	1	15.0	16.2	19.4	19.6	19.7	19.8	24.2	21.9
7	0	2	2	2	16.3	16.7	19.1	15.6	22.6	18.2	23.3	20.4
3	1	0	1	2	18.3	17.4	18.9	18.6	21.0	18.9	23.2	24.7
2	1	1	2	0	19.7	18.6	19.4	25.1	25.6	21.4	27.5	25.3
5	1	2	0	1	16.2	16.3	20.0	19.8	14.7	19.6	22.5	24.7
8	2	0	2	1	16.4	19.1	18.4	23.6	16.8	18.6	24.3	21.6
6	2	1	0	2	14.2	15.6	15.1	16.8	17.8	19.6	23.2	24.4
9	2	2	1	0	16.1	19.9	19.3	17.3	23.1	22.7	22.6	28.6

and others have argued that the use of signal to noise ratios, as advocated by Taguchi, is an inferior form of analysis than provided by classical analysis of variance. These criticisms seem to be well founded, but it should not be forgotten that they are matters of secondary importance. Taguchi got people to look at an important problem that had long been overlooked, variability. Improving on the details of analysis is the easy part. And even this easy part is far from being completed.

One thing Taguchi emphasized was the idea that in producing items designed to meet some target level, quality falls off as a function of the distance from the target. Thus, the goal of the process is, not only to have a process that is on target on average, but to have a process that in addition has as little variability as possible. The basic idea is that products should be designed so that they are robust to the conditions under which they have to operate. The procedure is called *robust parameter design*.

In Section 1 we will examine an experiment that focuses on variability. Note that the Hare data of Section 10.1 was also focused on reducing variability. In Section 2 we present Taguchi's signal to noise ratios and in Section 3 we present his form of analysis. Section 4 examines the role of noise factors, i.e., factors that contribute to the variability of the product but that can only be brought under control in exceptional circumstances. For example, how well a printer feeds paper may depend on the atmospheric humidity. With great effort, one can control the humidity in a room in order to examine the effect of humidity in an experiment. However, in practice, the printer will be used in rooms that do not have controlled humidity. The idea is to develop a printer that will be as insensitive as possible to humidity. In order to do that, one sets the printer up using variables that can be easily controlled, so as to make the printer robust to changes in humidity.

11.1 Experiment on Variability

Byrne and Taguchi (1989) and Lucas (1994) considered an experiment on the force y, measured in pounds, needed to pull tubing from a connector. Large values of y are good. The *controllable factors* in the experiment are as follows:

A: Interference (Low, Medium, High),

B: Wall Thickness (Thin, Medium, Thick),

C: Insertion Depth (Shallow, Medium, Deep),

D: Percent Adhesive (Low, Medium, High).

The treatment with, let's say, low interference, medium wall thickness, shallow insertion depth, and high percent adhesive will be denoted $a_0b_1c_0d_2$ where the letters identify the factors and the subscripts identify the factor levels. (The theory of such designs makes it convenient to start the subscripts at 0 but other authors use other methods for denoting treatments.) The data are given in Table 11.1 with the treatment subscripts in columns 2 through 5. Because we are interested in variability, similar to the Hare data, there are 8 observations on each of the observed treatments.

The choice of treatments is Taguchi's L_9 design. (The L_9 is equivalent to a Greco-Latin square.)

Taguchi refers to the observed treatment combinations as the *inner array*. Taguchi refers to the 8 observations on each treatment as observations on the *outer array*. Details of the outer array are discussed in Section 4.

The L_9 design is a fractional factorial involving four factors each at three levels. The experiment involves $3\times 3\times 3\times 3 = 3^4 = 81$ factor combinations of which only 9 are observed. With $3^4 = 81$ factor combinations and only 9 observed treatments, we have an $81/9 = 1/9$th replication. Briefly, it is a $1/9$th rep of a 3^4 design, sometimes written as a 3^{4-2} design. (Recall $1/9 = 3^{-2}$.) To define groups of treatments in the previous chapter for 2^f designs, we looked at adding some of the subscripts that define treatments and evaluating whether those sums were even or odd. Similarly, to define groups in this 3^4 design, we again add various subscripts to define treatment groups but now things get more complicated, cf. *TiD*, Chapter 3. Nonetheless, looking at the treatment subscripts given in the second through fifth columns of Table 11.1, we see that adding the b, c, and d subscripts always gives a multiple of 3. Similarly, the sum of the a and b subscripts plus *twice* the c subscript also is always a multiple of 3.

Fractional replications for 3^f factorials are discussed in *TiD*. For now it suffices to note that this L_9 design allows one to estimate all of the main effects. In fact, it only allows estimation of the main effects and it assumes the absence of any interactions. This is characteristic of the L_q designs suggested by Taguchi. He was only interested in main effects involving the control(able) factors.

11.2 Signal-to-Noise Ratios

Taguchi proposed the analysis of *signal-to-noise ratios* computed for each inner array treatment combination from its outer array data.

There are eight observations for each inner array factor combination in Table 11.1. These observations constitute the outer array data and will be discussed in Section 4. In the Taguchi analysis the outer array observations are treated as a random sample, even though, it turns out that they were obtained in a systematic fashion. In his analysis, the multiple observations in the outer array are summarized prior to analysis. The obvious summarization is the sample mean but additional common summary measures include the sample variance, sample standard deviation, or the logarithms of those values.

One of Taguchi's most controversial ideas was to summarize the outer array data using "signal-to-noise ratios." The idea is to maximize the appropriate signal-to-noise ratio. Let the observations be y_{io} with i — inner array and o — outer array, $o = 1,\ldots,N$. For minimizing a response y his signal-to-noise ratio is defined as

$$SN_{\min,i} \equiv -\log\left(\sum_{o=1}^{N} y_{io}^2/N\right) = -\log\left(\bar{y}_{i\cdot}^2 + \frac{N-1}{N}s_i^2\right).$$

Because of the minus sign, to make this large you need both $\bar{y}_{i\cdot}^2$ and s_i^2 to be small but of course there are many other functions of $\bar{y}_{i\cdot}$ and s_i^2 that could accomplish the same thing. For maximizing a response his signal-to-noise ratio is

$$SN_{\max,i} \equiv -\log\left(\sum_{o=1}^{N} 1/y_{io}^2 N\right),$$

apparently because maximizing y is the same as minimizing $1/y$ (for positive y). For minimizing variability around a target his signal-to-noise ratio is

$$SN_{\text{tar},i} \equiv \log\left(\frac{\bar{y}_{i\cdot}^2}{s_i^2}\right) = \log(\bar{y}_{i\cdot}^2) - \log(s_i^2).$$

The apparent rationale is that if $\bar{y}_{i\cdot}$ always remains close to the target, you just want to minimize s_i^2.

Table 11.2: *Bryne-Taguchi* 3^{4-2} *Design*

Run					Outer Array Summaries		
i	a	b	c	d	$\bar{y}_{i\cdot}$	s_i^2	$SN_{\max,i}$
1	0	0	0	0	17.5250	13.050714	5.532040
4	0	1	1	1	19.4750	8.447857	5.876575
7	0	2	2	2	19.0250	8.313571	5.833544
3	1	0	1	2	20.1250	6.747857	5.964675
2	1	1	2	0	22.8250	11.747857	6.195688
5	1	2	0	1	19.2250	11.422143	5.831468
8	2	0	2	1	19.8500	8.908571	5.920132
6	2	1	0	2	18.3375	14.248393	5.717851
9	2	2	1	0	21.2000	15.585714	6.021715

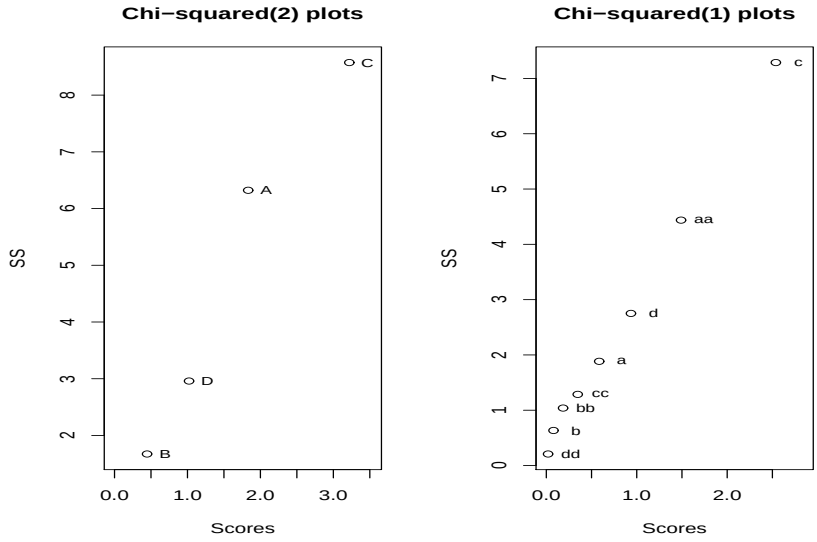

Figure 11.1: χ^2 *plots for main effects: Means of outer array.*

An apparently common approach is to divide the control factors into two groups. First identify control factors that affect the signal-to-noise ratio and use them to maximize it. Then use the control factors that do not have much affect on the SN ratio to try to put the process on target.

11.3 Taguchi Analysis

This section is just copied from *TiD* with corrections to table and figure numbers.

Table 11.2 contains three summary statistics to be used in a Taguchi analysis of the Byrne-Taguchi data. We fit a main-effects model to each of $\bar{y}_{i\cdot}$, $\log(s_i)$, and $SN_{\max,i}$. For each dependent variable we constructed two χ^2 plots. The first is a $\chi^2(2)$ plot for the main-effect sums of squares. The second plot is a $\chi^2(1)$ plot based on treating the factor subscripts (associated with ordered levels) as regression variables and fitting quadratic polynomials in the main effects. (The quadratic effects really just measure nonlinearity.) This gives sums of squares for a linear (e.g. a) and quadratic (e.g. aa) contrast in each main effect. Figures 11.1, 11.2, and 11.3 contain the χ^2 plots for the different dependent variables.

I don't see any clear evidence for the existence of main effects in either the mean or the log-standard-deviation plot. But I can imagine someone else arguing that all or nearly all of the effects

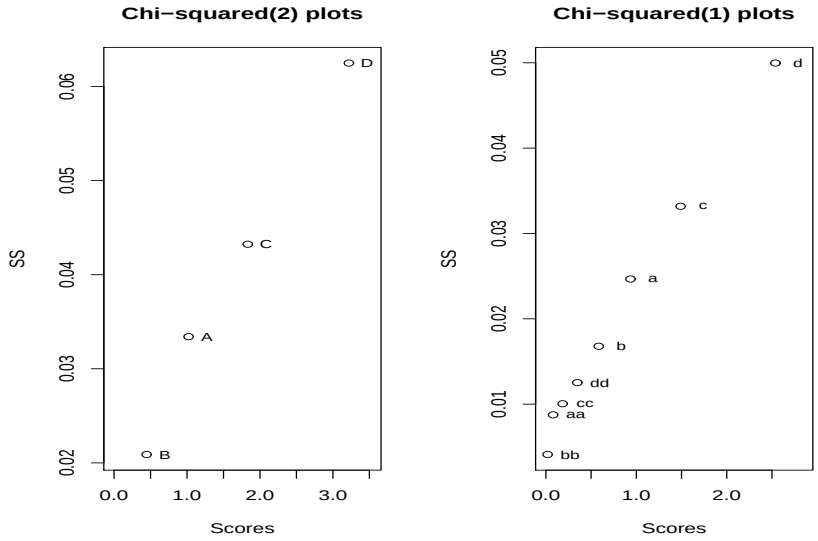

Figure 11.2: χ^2 *plots for main effects: Log standard deviations of outer array.*

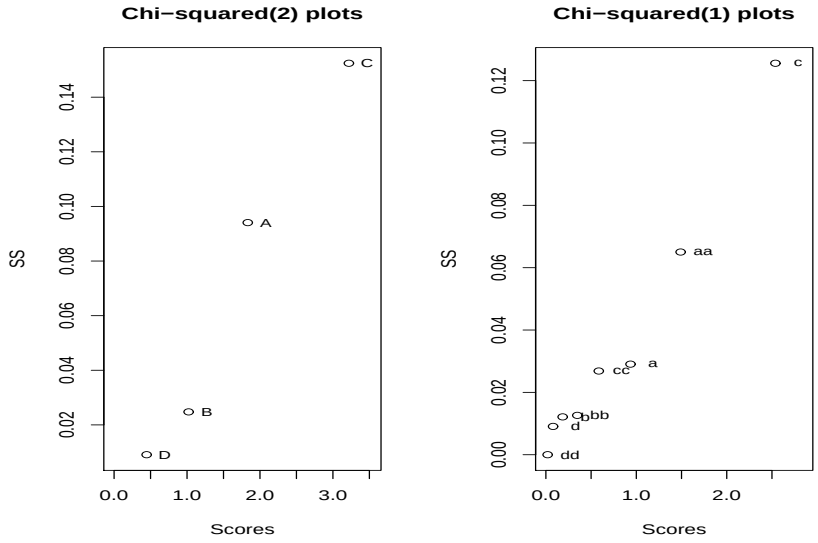

Figure 11.3: χ^2 *plots for main effects:* $SN_{\max}$ *of outer array.*

are important. For the signal-to-noise ratio I again see no *clear* evidence of effects but some evidence for one or possibly two contrasts having effects. The two largest sums of squares are for the linear effect in C and the quadratic effect in A.

To make sense of any important effects we would look at means plots. These are given in Figures 11.4, 11.5, and 11.6. We will not discuss the means plots for the means or log standard deviations of the outer array data because they displayed no obvious main effects. For the signal-to-noise ratio the two largest $\chi^2(1)$ values were for the curvature in A and the linear effect in C. Since interest is in maximizing the signal-to-noise ratio, the recommendation would be to pick the middle level of A and, despite the importance of the linear effect in C (which really only looks at the difference

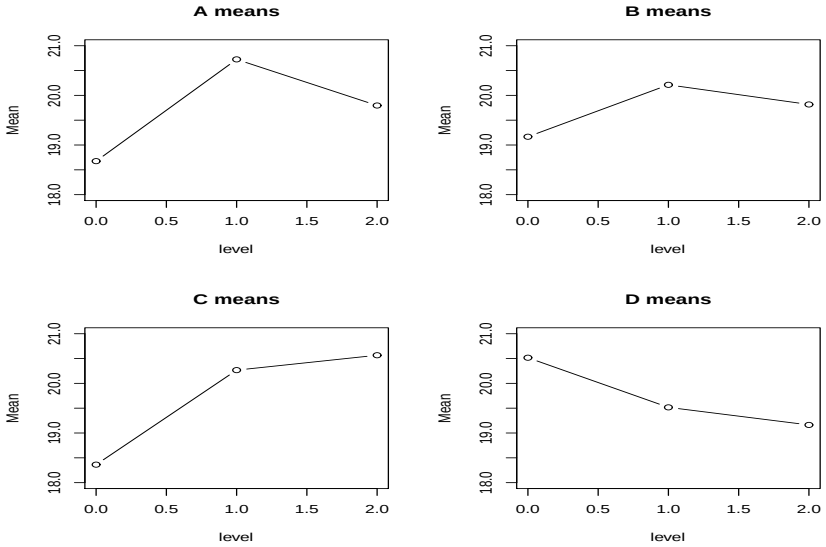

Figure 11.4: *Means plots for main effects: Means of Taguchi outer array.*

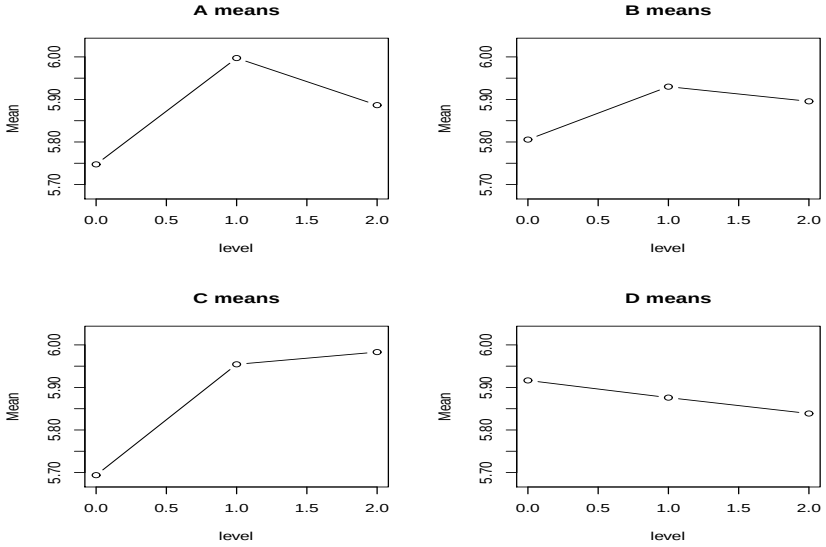

Figure 11.5: *Means plots for main effects: Log standard deviations of Taguchi outer array.*

between the low and high levels), it looks like either the high or middle level of C should work reasonably well.

In practice you have to play off the experimental results against the production costs of various techniques. For example, if two levels have roughly the same effects, obviously you would choose the more inexpensive level. If two levels have radically different costs, it is harder to decide whether improved performance is worth the cost.

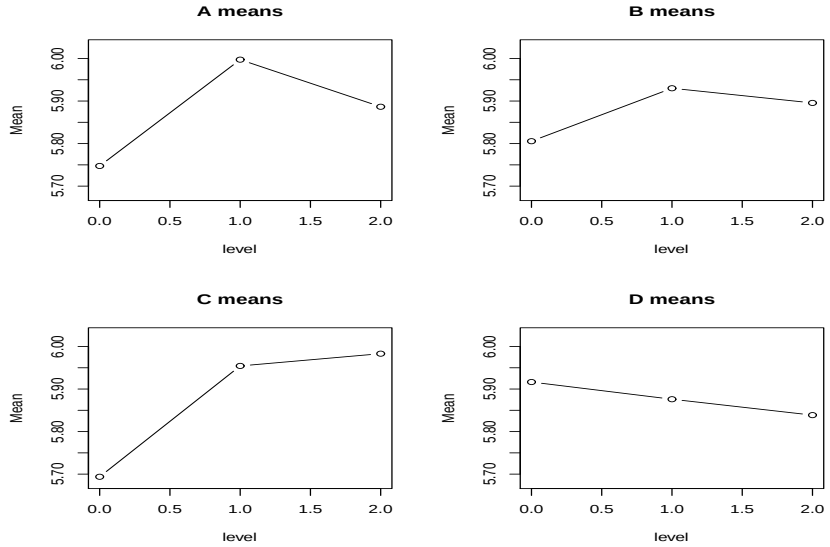

Figure 11.6: *Means plots for main effects:* $SN_{\max}$ *of Taguchi outer array.*

11.4 Outer Arrays

In Section 1 we examined a fractional replication on which several observations were taken for each treatment combination that appeared in the experimental design. The several observations were used to estimate variability as though they were from a random sample. Taguchi actually suggested another approach rather than random sampling. In the tubing experiment, it was thought that most of the variability would be due to three *noise* factors: the condition time, the condition temperature, and the condition relative humidity. In the field or in a manufacturing plant, these are not conditions that can readily be controlled. However, for the purposes of designing the product, we could go to the trouble of controlling them. Rather than allowing these noise factors to vary randomly, the levels of these noise factors were actually set by design. The idea is that there is some mean and standard deviation for these noise factors and that the selected values are one standard deviation from the mean.

The complete set of factors and levels are given below.

A: Interference (Low, Medium, High),

B: Wall Thickness (Thin, Medium, Thick),

C: Insertion Depth (Shallow, Medium, Deep),

D: Percent Adhesive (Low, Medium, High).

E: Condition Time (24hr, 120hr)

F: Condition Temperature (72°F, 150°F)

G: Condition Relative Humidity (25%, 75%)

The data with all of the factors identified are presented in Table 11.2. One approach to analyzing this data is to ignore the information about the noise factors and just perform the analysis of Section 3. While it would be interesting to see, the 7 factor ANOVA for these data is beyond the introduction/survey ambitions of this book. Such an analysis would probably focus on low-order interactions between the control factor main effects and outer array effects looking for control factor levels for which the response changes little as the noise factors change.

In general, Taguchi divides the factors of interest into control factors, those that can be readily

Table 11.3: *Taguchi* $3 \times 3 \times 3 \times 3 \times 2 \times 2 \times 2$ *design.*

Run	a	b	c	d	$e_0f_0g_0$	$e_0f_0g_1$	$e_0f_1g_0$	$e_0f_1g_1$	$e_1f_0g_0$	$e_1f_0g_1$	$e_1f_1g_0$	$e_1f_1g_1$
1	0	0	0	0	15.6	9.5	16.9	19.9	19.6	19.6	20.0	19.1
4	0	1	1	1	15.0	16.2	19.4	19.6	19.7	19.8	24.2	21.9
7	0	2	2	2	16.3	16.7	19.1	15.6	22.6	18.2	23.3	20.4
3	1	0	1	2	18.3	17.4	18.9	18.6	21.0	18.9	23.2	24.7
2	1	1	2	0	19.7	18.6	19.4	25.1	25.6	21.4	27.5	25.3
5	1	2	0	1	16.2	16.3	20.0	19.8	14.7	19.6	22.5	24.7
8	2	0	2	1	16.4	19.1	18.4	23.6	16.8	18.6	24.3	21.6
6	2	1	0	2	14.2	15.6	15.1	16.8	17.8	19.6	23.2	24.4
9	2	2	1	0	16.1	19.9	19.3	17.3	23.1	22.7	22.6	28.6

controled, and noise factors, those that cannot readily be controlled (but can be controlled for the purposes of the experiment). Taguchi assumes, or perhaps more properly defines, control factors to have no interactions with each other. This has been a major source of criticism for Taguchi's methods. The overall idea is to use interactions between the control factors and the noise factors to reduce variability in the end product.

The basic experimental designs suggested by Taguchi involve a highly fractionated design on the control factors, one that allows estimation of all main effects (and usually only main effects), a separate design on the noise factors, and then to observe every treatment from the control factor design with every treatment from the noise factor design. In our example, we have a 1/9th rep of a 3^4 design for the control factors and a full replication of a 2^3 for the noise factors. This process of crossing individual designs is often wasteful of experimental material. Setting up an appropriate fractional replication for the entire collection of factors is more economic, cf Shoemaker et al. (1991). It is also often suggested that there is little need to look at three levels of the control factors, two levels will usually be enough. (Taguchi used both 2 and 3 level designs.)

A correct analysis of the data depends crucially on how the outer array data are physically collected. If you really wanted to get a random sample for the outer array you should list each of the 9 treatments 8 times to create a list of 72 treatments and run them in random order without trying to control the noise factors. That would justify the analysis conducted in Section 3. With the control factor – noise factor paradigm, there will be a real temptation to run the experiment as a split plot experiment, cf. Christensen (1996, 2015). When the 8 observations on each inner array treatment correspond to a different outer array treatment, to avoid a split plot experiment you would again create a list of 72 total treatments and run them in random order. This seems like it might frequently be inconvenient.

There are two other ways you could run the experiment that both result in a split plot experiment. You could fix the outer array conditions (which are harder to fix) and then run each of the 9 inner array treatments. (Relative to a split plot the outer array becomes the whole plot treatments and the inner array becomes subplot treatments.) This is good for looking at inner array main effects and their interactions with the outer array. The appropriate error is some work to find. The Taguchi analysis of Section 3 that just ignores the outer array treatments is more consistent with a different split plot design in which you fix an inner array treatment and then take 8 observations, one for each outer array condition. If the outer array conditions are harder to fix, this seems much less attractive. Of course the outer array conditions may be impossible to fix in manufacturing but may be easy to fix in a pilot plant. It depends on individual circumstances. The proper analysis of split plot designs is discussed in Christensen (1996, 2015).

The reader should be aware that seriously misleading results can be obtained when a split plot design is not analyzed properly. However, that is largely due to the creation of inappropriate estimates of error. In the Taguchi set-up, there typically are no estimates of error and χ^2 plots look for odd behavior among main effect terms that should behave similarly if there are no effects.

Taguchi seems to treat the outer array as a random sample of the noise conditions that one would

encounter. In fact it is a systematic sample of noise conditions. Although it seems mathematically unjustified, as a practical matter, it seems to me eminently reasonable to treat the outer array as random (because it is designed to be representative of the noise conditions).

11.5 Discussion

It has been suggested that with a good design, the form a analysis may be of little importance. Perhaps just pick the design point that has the best characteristics and use that. Our examination of Hare's experiment suggests that is not a good strategy. In Hare's experiment, our suggested action was to change to a 7 day delay. However, the current operating setup but with a change to a 7 day delay was NOT one of the design points in the 1/2 rep., so picking the best of the observed treatments would lead to a different result.

Ultimately the goal of experiments is to better understand (model in the least technical sense) the process. With better understanding may come leaps of improvement rather than evolutionary improvements. The Shewhart cycle should be used in experimentation. *Evolutionary Operation (EVOP)* is a formal version of such a process, cf. Box and Draper (1969).

11.6 Exercises

EXERCISE 11.6.1. Investigate the data in Table 11.4 from Luner (1994, qe94-698). With 3 observations the outer array can hardly be anything but either a random sample of observations or three levels of one noise factor. The inner array is a nonstandard *central composite design* with replications at the center point. The first 8 observations are are 2^3 full factorial without replications. With $f = 3$, the first 8 observations under this ± 1 notation have (A,B,C) scores satisfying $f = 3 = (\pm 1)^2 + (\pm 1)^2 + (\pm 1)^2$ which is their squared length from the origin (0,0,0) of three-dimensional space. The next 6 observations are a star design that looks to be incorrectly constructed. Normally, star points would be $(\pm\sqrt{f}, 0, 0)$, $(0, \pm\sqrt{f}, 0)$, $(0, 0, \pm\sqrt{f})$ so that they have squared length $f = (\pm\sqrt{f})^2 + (0)^2 + (0)^2$ etc., but here $\sqrt{3} \neq 1.682$. The last 6 design observations are true replications at the center (origin) and provide an estimate of pure error. Central composite designs are classically used for fitting quadratic response surfaces.

Does the pairing of the replications seem suspicious? Runs 14 and 15 occur as consecutive design points as do 19 and 18, 6 and 5, and 3 and 2.

EXERCISE 11.6.2. Examine the data of Moen et al. (1991) and Bisgaard and Fuller (1996, qe96-373) given in Table 11.5 on a Solenoid experiment. The factors and levels are:
A: length of armature (.595in or .605in),
S: spring load (70g or 100g),
B: bobbin depth (1.095in or 1.105in),
T: length of tube (.50in or .51in).
The existence of an "outer array" is implicit from the dependent variable.

Table 11.4: *Taguchi: qe94-698*

	Inner Array			Outer Array		
Run	A	B	C	1	2	3
20	−1	−1	−1	39	34	42
9	−1	−1	1	80	71	91
11	−1	1	−1	52	44	45
14	−1	1	1	97	68	60
15	1	−1	−1	60	53	68
10	1	−1	1	113	104	127
13	1	1	−1	78	64	65
1	1	1	1	130	79	75
7	−1.682	0	0	59	51	60
4	1.682	0	0	115	102	117
19	0	−1.682	0	50	43	57
18	0	1.682	0	88	49	43
12	0	0	−1.682	54	50	60
8	0	0	1.682	122	109	119
6	0	0	0	87	78	89
5	0	0	0	86	79	85
17	0	0	0	88	81	87
3	0	0	0	89	82	87
2	0	0	0	86	79	88
16	0	0	0	88	79	90

Table 11.5: *Solenoid Experiment: qe96-373*

A	S	B	T	$-\log(s^2)$
0	0	0	0	6.44
1	0	0	0	3.67
0	1	0	0	7.82
1	1	0	0	4.61
0	0	1	0	7.82
1	0	1	0	9.21
0	1	1	0	5.99
1	1	1	0	9.21
0	0	0	1	6.44
1	0	0	1	3.79
0	1	0	1	5.32
1	1	0	1	5.63
0	0	1	1	7.82
1	0	1	1	7.82
0	1	1	1	7.82
1	1	1	1	9.21

11.7 Ideas on Analysis

This section is obviously far from complete and may get eliminated.

plot s_i versus $\bar{y}_i$

if equal variances, analyze $\bar{y}_i$s if unequal variances, analyze $\log(s_i)$ as well as transform data before doing analysis of $\bar{y}_i$s . Use Box-Cox.

$x = (x_1, x_2)$

$$y_i = \mu(x_i) + \varepsilon_i$$

$\sigma^2_{h(y)} \doteq [h'(\mu(x))]^2 \sigma^2_y(x)$

minimize $\mathrm{E}(y-T)^2 = \sigma^2_y(x) + [\mu(x) - T]^2$.

by choosing x_1 appropriately, change to minimize

$$\mathrm{E}(y-T)^2 = \sigma^2_y(x_1) + [\mu(x) - T]^2 \doteq \sigma^2_{h(y)}/[h'(\mu(x))]^2 + [\mu(x) - T]^2$$

Pick x_1 to minimize $\sigma^2_y(x_1) \doteq \sigma^2_{h(y)}/[h'(\mu(x))]^2$.

Pick x_2 to achieve the target, i.e., minimize $\sigma^2_y(x_1) \doteq \sigma^2_{h(y)}/[h'(\mu(x))]^2$. The min occurs at μ satisfying

$$\mu = \left\{ T + \frac{3}{2}[h'(T)]^{-3} h_{\prime\prime}(T) \sigma^2_{h(y)} \right\}$$

If $h(y) = \log(y)$, $\mu \doteq \log(T) - \frac{3}{2}\sigma^2_{h(y)}$.

If $h(y)$ is a variance stabilizing transformation, $\sigma^2_{h(y)}$ does not depend on x, so all of x (not just x_2) can be used to minimize the function.

11.7.1 A modeling approach

Model

$$\hat{y} = f(x) + g_1(x) z_1 + g_2(x) z_2$$

To set target, use

$$\hat{E}(y) = f(x)$$

To minimize variance use

$$\hat{\sigma}^2_y = MSPE + [g_1(x)]^2 + [g_2(x)]^2$$

Suppose we can choose x_1 so that

$$\hat{\sigma}^2_y = MSPE + [g_1(x_1)]^2 + [g_2(x_1)]^2$$

Once again, pick x_1 to minimize variance and then pick x_2 to get on target (or to make response large or small).

Assuming that noise factors are uncorrelated and variance 1 on scale used. (i.e., chosen at ± 1 std dev from mean).

Appendix A: Multivariate Control Charts

Multivariate data involves taking several measurements $y_1, \ldots, y_q$ on each item. With multivariate data we can and should control chart the data on each variable but we can also combine all the variables for each item into multivariate control charts. Sometimes these will catch things that are not apparent in the univariate charts. Multivariate charts are based on squared Mahalanobis distances for looking at means and generalized variances for looking at dispersions. Discussion of generalized variances is deferred to Section 4.

EXAMPLE A.0.1. Table A.1 contains the Coleman Report data (Christensen, 2015, Chapter 6 or Christensen, 1996, Chapter 7) on some Northeastern US schools. The variables are x—socioeconomic status, y—reading ability. Case 17 has been modified to be an outlier (special cause). Despite the fact that the new values for case 17 are not unusual in either the x or y coordinates, the scatter plot in Figure A.1 shows how weird the new case 17 looks relative to the other data. With more that $q = 2$ variables there is no assurance that such points would be apparent from doing two-dimensional scatterplots. □

Using the matrix tools of Section 5.7 denote an arbitrary collection (vector) of measurements $y = (y_1, \ldots, y_q)'$, with the measurements on an ith individual denoted $y_i = (y_{i1}, \ldots, y_{iq})'$. When collecting data in rational subgroups, the jth individual in the ith group has measurements $y_{ij} = (y_{ij1}, \ldots, y_{ijq})'$.

For q-dimensional vectors of multivariate observations y with $\mathrm{E}(y) = \mu$ and $\mathrm{Cov}(y) = \Sigma$, the theoretical squared *Mahalanobis distance* from y to the center of the distribution is

$$D^2 \equiv (y-\mu)'\Sigma^{-1}(y-\mu).$$

Using Christensen (2020, Theorem 1.6.1) it is easy to see that

$$\mathrm{E}(D^2) = \mathrm{E}[(y-\mu)'\Sigma^{-1}(y-\mu)] = q,$$

so q could be the center of a D^2 control chart. Computing the control limits requires computing the variance and the variance computation assumes that the observations are (multivariate) normal. Using a well-known variance formula, e.g. Christensen (2019, Theorem 4.6.1), it is also easy to see that

$$\mathrm{Var}(D^2) = \mathrm{Var}[(y-\mu)'\Sigma^{-1}(y-\mu)] = 2q,$$

Table A.1: *Modified Coleman Report data. Case 17 modified: New(True).*

School	y	x	School	y	x
1	37.01	7.20	11	23.30	−12.86
2	26.51	−11.71	12	35.20	0.92
3	36.51	12.32	13	34.90	4.77
4	40.70	14.28	14	33.10	−0.96
5	37.10	6.31	15	22.70	−16.04
6	33.90	6.16	16	39.70	10.62
7	41.80	12.70	17	25 (31.80)	11 (2.66)
8	33.40	−0.17	18	31.70	−10.99
9	41.01	9.85	19	43.10	15.03
10	37.20	−0.05	20	41.01	12.77

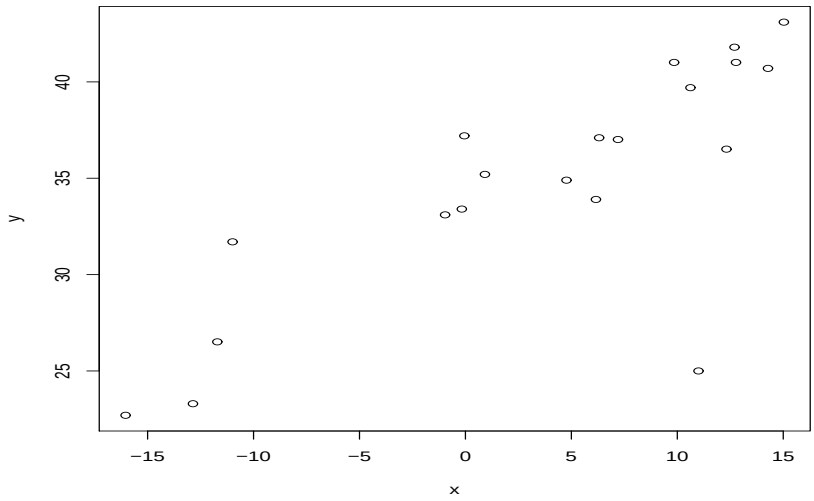

Figure A.1: *Scatter plot of Coleman report school data:. New case 17 in bottom right.*

so the control limits on a D^2 chart could be set at $q \pm 3\sqrt{2q}$. The lower control limit can be ignored (is not positive) unless $q > 18$. Of course this all requires us to know μ and Σ, assumptions we will soon relax, and typically more sophisticated procedures are used.

A.1 Statistical Testing Background

Standard univariate charts for individuals or means are not directly based on corresponding statistical tests. Standard multivariate charts for individuals and means, however, correspond closely to multivariate statistical tests. Before discussing how multivariate charts are defined from multivariate statistical tests, we illustrate the correspondence between univariate charts and statistical tests. Multivariate tests are discussed generally in Christensen (1996, 2015, 2019 – from easiest to hardest discussions).

For univariate grouped data y_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,N$, a means chart is out of control if $\bar{y}_{i\cdot}$ is outside the control limits $\bar{y}_{\cdot\cdot} \pm 3\sqrt{MSE/N}$. Equivalently, it is out of control if $|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}| > 3\sqrt{MSE/N}$ or if

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|}{\sqrt{MSE/N}} > 3. \tag{A.1.1}$$

From the statistical theory of testing contrasts in a one-way analysis of variance with normal distributions, one can show (see Christensen, 1996, particularly the discussions of contrasts and Ott's Analysis of Means) that an α level statistical test of whether the group i population mean is different from the average of the other group means is rejected if

$$\frac{|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}|}{\sqrt{MSE\frac{(n-1)/n}{N}}} > t(1-\alpha, n(N-1)) = \sqrt{F(1-\alpha,1,n(N-1))}. \tag{A.1.2}$$

Relative to inequality (A.1.1), the multiplier 3 from the control chart has been replaced by a percentile of a t distribution and there is an additional multiplier of $\sqrt{(n-1)/n}$ in the denominator. Evaluating inequality (A.1.2) is equivalent to rejecting when $\bar{y}_{i\cdot}$ is outside the test limits

$$\bar{y}_{\cdot\cdot} \pm \sqrt{F(1-\alpha,1,n(N-1))}\sqrt{MSE(n-1)/nN}.$$

Rather than using inequality (A.1.2) directly, multivariate means charts square both sides of the inequality, so they evaluate whether

$$\frac{(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{MSE\frac{(n-1)/n}{N}}=\left(\frac{|\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}|}{\sqrt{MSE\frac{(n-1)/n}{N}}}\right)^2>[t(1-\alpha,n(N-1))]^2=F(1-\alpha,1,n(N-1))$$

or whether

$$\frac{N(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{MSE}>\frac{n-1}{n}F(1-\alpha,1,n(N-1)). \tag{A.1.3}$$

In fact, when $q=1$, the multivariate means chart, as typically used, reduces exactly to determining whether inequality (A.1.3) is satisfied.

We discussed earlier why viewing control charts as statistical tests was inappropriate because control charts are designed not as tests but as operational definitions.

A.2 Multivariate Individuals Charts

From a sample of n multivariate observations y_i we can compute the sample mean vector $\bar{y}_{\cdot}$ and the sample covariance matrix

$$S\equiv\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y}_{\cdot})(y_i-\bar{y}_{\cdot})'.$$

Alternatively, one could (and perhaps should) use the covariance matrix estimate based on running differences (moving ranges)

$$\tilde{S}\equiv\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_{i+1}-y_i)(y_{i+1}-y_i)'$$

which, like S, can be shown to have

$$\mathrm{E}\left(\tilde{S}\right)=\Sigma.$$

As with its univariate version discussed in Section 4.1, $\tilde{S}$ is subject to far less bias than is S when the y_i process is subjected to a mean shift at some time (or times).

Compute the estimated squared Mahalanobis distances for each observation as, say,

$$\tilde{D}_i^2\equiv(y_i-\bar{y}_{\cdot})'\tilde{S}^{-1}(y_i-\bar{y}_{\cdot}).$$

Again an individuals chart can be constructed with control limits $q\pm3\sqrt{2q}$. To test whether the process is on target, i.e. $H_0:\mu=\mu_0$, plot

$$\tilde{D}_{i0}^2\equiv(y_i-\mu_0)'\tilde{S}^{-1}(y_i-\mu_0)$$

with the same control limits.

The more commonly used multivariate individuals chart is based on statistical testing. Sullivan and Woodhall (1996) review the relevant literature and suggest an upper control limit that uses percentiles of an approximate Beta distribution.

$$UCL=\frac{(n-1)^2}{n}\mathrm{Beta}(1-\alpha,q/2,(f-q-1)/2)\qquad\text{where}\qquad f=\frac{1}{2}\left(\frac{2(n-1)^2}{3n-4}\right).$$

The distribution theory is somewhat more complex than it is for means charts. Minitab uses $\alpha=0.00134989803156746$. A center line can be defined using $\alpha=0.5$ and the lower control limit is 0. When using the sample covariance S instead of $\tilde{S}$, an exact Beta distribution applies,

$$UCL=\frac{(n-1)^2}{n}\mathrm{Beta}(1-\alpha,q/2,(n-q-1)/2).$$

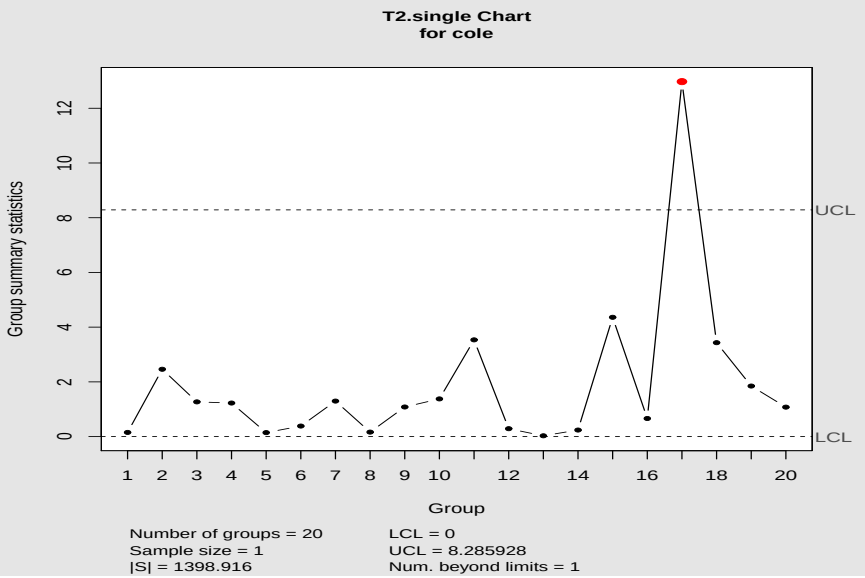

Figure A.2: *Individuals "T^2" chart of school scores.*

As a statistical testing procedure (as opposed to an operation definition) the issues are further complicated by the fact that (nonindependent) tests are being executed for each 1 from 1 to n so the overall error rate for finding a special cause when none exists is complicated.

EXAMPLE A.2.1. Figure A.2 shows an individuals chart for the Coleman Report data with the new case 17 having by far the largest $\hat{D}_i$ value. The index for School is used in place of time. This is an `mqcc` plot using R and I think it uses S. I believe the Minitab default is to use $\tilde{S}$. □

R code for Figures A.1 and A.2 follows. The code also produces an ellipse chart that puts an ellipse around most of the data in Figure A.1 but excludes point 17.

```
# Read the data
coleman.slr <- read.table("C:\\E-drive\\Books\\ANREG2\\NewData\\tab6-1.dat",
  sep="",col.names=c("School","x","y"))
attach(coleman.slr)
coleman.slr
#summary(coleman.slr)
x[17]=11
y[17]=25
plot(x,y)
cole=data.frame(x,y)
cole

#install.packages("qcc")
library(qcc)
clr=mqcc(cole, type ="T2.single")
ellipseChart(clr)

# The following code is for the generalized variance chart but also shows
# alternatives to qcc's mqcc program.
#install.packages("MSQC")
```

```
library(MSQC)
gen.var(cole)
clr1=mult.chart(type = c("chi", "t2", "mewma", "mcusum", "mcusum2"))
```

A.3 Multivariate Means (T^2) Charts

Prior to the introduction of Mahalanobis distance, Hotelling had introduced an overall test for whether a random sample of size n was on target at μ_0. His test statistic was

$$T^2 = n(\bar{y}_{\cdot} - \mu_0)' S^{-1} (\bar{y}_{\cdot} - \mu_0),$$

which is the squared Mahalanobis distance from $\bar{y}_{\cdot}$ to the hypothesized center of its distribution. (Recall that $[\mathrm{Cov}(\bar{y}_{\cdot})]^{-1} = n\Sigma^{-1}$.) Hotelling also related the distribution of T^2 to an F distribution when the data y_i are multivariate normal. Hotelling's testing idea has been enlarged to find the usual upper control limit for a multivariate means chart. Despite such charts being motivated by estimated Mahalanobis distances rather than Hotelling's T^2, the charts are often referred to as T^2 charts as they are closely, but not immediately, related to Hotelling's statistic.

When using n rational subgroups of size N to obtain control data y_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,N$, for each subgroup i we compute the subgroup mean vector $\bar{y}_{i\cdot}$ and the subgroup covariance matrix S_i. Compute the grand mean

$$\bar{y}_{\cdot\cdot} = (\bar{y}_{1\cdot} + \cdots + \bar{y}_{n\cdot})/n,$$

the pooled covariance matrix

$$S_p = (S_1 + \cdots + S_n)/n,$$

and the group-wise estimated squared Mahalanobis distances

$$\hat{D}_i^2 = N(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})' S_p^{-1} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

Again this could be related to the control limits $q \pm 3\sqrt{2q}$ but we can do something more statistically precise because $\hat{D}_i^2$ is closely related to the Lawley-Hotelling trace statistic T_i^2 for testing whether group i has the same mean as the (average of the) other groups. Similar to the computation in Section 1, the test statistic reduces to

$$T_i^2 = \frac{n}{n-1}\hat{D}_i^2.$$

When testing a single (rank 1, estimable, multivariate) linear hypothesis (like we are doing) it is quite well known that

$$\frac{dfE - q + 1}{dfE\ q} T^2 \sim F(q, dfE - q + 1).$$

For our rational subgroup data structure $dfE = n(N-1)$. Some simple algebra shows that

$$\frac{nN - n - q + 1}{n(N-1)\ q} \frac{n}{n-1} \hat{D}_i^2 \sim F(q, nN - n - q + 1)$$

or

$$\hat{D}_i^2 \sim \frac{(N-1)(n-1)q}{nN - n - q + 1} F(q, nN - n - q + 1).$$

The upper contral limit on a $\hat{D}_i^2$ chart can be, and typically is, taken as a multiple of a percentile of an F distribution, namely,

$$UCL \equiv \frac{(N-1)(n-1)q}{nN - n - q + 1} F(1 - \alpha, q, nN - n - q + 1)$$

for a small α. Again, for $q = 1$ this reduces exactly to the procedure embodied in (A.1.3). Minitab

uses $\alpha = 0.00134989803156746$. The lower control limit is 0. A center line is often given at $\frac{(N-1)(n-1)q}{nN-n-q+1}F(0.5,q,nN-n-q+1)$.

These procedures are for process analysis. Multivariate one-way ANOVA theory can provide a predictive test for process control. One can also perform multivariate EWMA and CUSUM charts by applying the procedures to the $\hat{D}_i^2$s.

A.4 Multivariate Dispersion Charts

One can chart each of the q individual variances (and the $q(q-1)/2$ covariances) but for multivariate charting purposes, covariance matrices are typically summarized by the *generalized variance* of a random vector. The generalized variance is just the determinant of the covariance matrix, say $|\Sigma|$. In $q = 1$ dimension, the generalized variance becomes just the variance. We primarily concern ourselves with data from rational subgroups. While constructing a generalized variance chart is analogous to constructing an s^2 chart, the actual construction is more in the spirit of constructing s charts due to the numerous bias corrections that are involved. For individual data vectors y_i, the *multivariate moving range chart* is an individuals chart on the values $(y_{i+1}-y_i)'(y_{i+1}-y_i)$.

For multivariate normal rational subgroup data, it is known that for some constants $b_{j,df}$ (given later),

$$\mathrm{E}(|S_i|) = b_{1,N-1}|\Sigma|, \qquad \mathrm{E}(|S_p|) = b_{1,n(N-1)}|\Sigma|$$

and

$$\mathrm{Var}(|S_i|) = b_{2,N-1}|\Sigma|^2, \qquad \mathrm{Var}(|S_p|) = b_{2,n(N-1)}|\Sigma|^2.$$

With no statistical testing involved, the theoretical control limits when plotting $|S_i|$ are

$$b_{1,N-1}|\Sigma| \pm 3\sqrt{b_{2,N-1}|\Sigma|^2}$$

or

$$|\Sigma|\left[b_{1,n(N-1)} \pm 3\sqrt{b_{2,n(N-1)}}\right].$$

If we estimate $|\Sigma|$ using $|S_p|/b_{2,n(N-1)}$ we get estimated control limits of

$$\frac{|S_p|}{b_{2,n(N-1)}}\left[b_{1,n(N-1)} \pm 3\sqrt{b_{2,n(N-1)}}\right].$$

Not that they are very interesting but

$$b_{1,df} \equiv \frac{1}{df^q}\prod_{k=0}^{q-1}(df-k)$$

and

$$b_{2,df} \equiv \frac{1}{df^{2q}}\prod_{k=0}^{q-1}(df-k)\left[\prod_{k=0}^{q-1}(df-k+2) - \prod_{k=0}^{q-1}(df-k)\right].$$

When $q = 1$, the generalized variance chart reduces to the s^2 chart presented in Section 4.2. In particular, $|S_i| = s_i^2$, $|\Sigma| = \sigma^2$, and $b_{1,N-1} = 1$, so

$$\mathrm{E}(s_i^2) = \mathrm{E}(|S_i|) = b_{1,N-1}|\Sigma| = \sigma^2.$$

Also, $b_{2,N-1} = 2/(N-1)$, so

$$\mathrm{Var}(s_i^2) = \mathrm{Var}(|S_i|) = b_{2,N-1}|\Sigma|^2 = 2\sigma^4/(N-1).$$

References

Aceves-Mijares, Mariano, Murphy-Arteaga, Roberto, Torres-Jacome, Alfonso, and Calleja-Arriaga, Wilfrido (1996). Quality assurance in polysilicon deposition using statistics. *Quality Engineering*, **8**, 255-262.

Anand, K. N. (1994). Improving paraffin wax yield through process optimization using Taguchi's method of experimentation. *Quality Engineering*, **6**, 39-56.

Aravamuthan, Raja and Yayin, Irmak (1994). Application of response surface methodology for the maximization of concora crush resistance of paperboard. *Quality Engineering*, **6**, 1-20.

Banks, David (1993). Is Industrial Statistics Out of Control? *Statistical Science*, Vol. **8**, 356-377.

Bisgaard, Soren and Fuller, Howard T. (1996). Reducing variation with two-level factorial experiments. *Quality Engineering*, **8**, 373- 377.

Box, George (1988). Signal-to-Noise Ratios, Performance Criteria, and Transformations. *Technometrics*, **30**, 1-17.

Box, George and Bisgaard, Soren (1994). Iterative analysis of data from two-level factorials. *Quality Engineering*, **6**, 319-330.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley and Sons, New York.

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation: A Statistical Method for Process Improvement*. John Wiley and Sons, New York.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, John Wiley and Sons, New York.

Box, George and Luceno, Alberto (1997). *Statistical Control by Monitoring and Feedback Adjustment*. John Wiley and Sons, New York.

Bullington, Robert G., Lovin, Scott, Miller, David M., and Woodall, William H. (1993). Improvement of an industrial thermodstat using designed experiments. *Journal of Quality Technology*, **25**, 262-270.

Byrne, D. M. and Taguchi, S. (1989). The Taguchi approach to parameter design. *Quality Progress*, **December**, 19-26.

Carter, Charles, W. (1996). Sequence-leveled experimental designs. Part III. Measurement and methodology. *Quality Engineering*, **8**, 499- 503.

Chao, M. T. and Cheng, Smiley W. (1996). Semicircle control chart for variables data. *Quality Engineering*, **8**, 441-446

Chapman, Robert E. (1996). Photochemistry multiple response co- optimization. *Quality Engineering*, **8**, 31-45.

Christensen, Ronald (1996). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, London. www.stat.unm.edu/~fletcher/anreg.pdf

Christensen, Ronald (2015). *Analysis of Variance, Design, and Regression: Linear Modeling of Unbalanced Data*, Second Edition. Chapman and Hall, London.

Christensen, Ronald (2020). *Plane Answers to Complex Questions: The Theory of Linear Models*, Fifth Edition. Springer-Verlag, New York.

Christensen, Ronald (2019). *Advanced linear Modeling: Statistical Learning and Dependent Data*, Third Edition. Springer-Verlag, New York.

Christensen, Ronald (2017). *Topics in Experimental Design*. www.stat.unm.edu/~fletcher/TopicsInDesign.

Christensen, R., and Huzurbazar, A. V. (1996). A note on augmenting resolution III designs. *The American Statistician*, **50**, 175-177.

Clark, Jill M. and Milligan, Glenn (1994). How sweet it is — quality management in a "Honey House": The sticky quality of problems of honey. *Quality Engineering*, **6**, 379-399.

Collins, William H. and Collins, Carol B. (1994). Including residual analysis in designed experiments: case studies. *Quality Engineering*, **6**, 547-565

Cox, D. R. (1958). *Planning of Experiments*. John Wiley and Sons, New York. (One of my favorite books.)

Deming, W. E. (1986). *Out of the Crisis*. MIT Center for Advanced Engineering Study, Cambridge, MA. (His classic work.)

Deming, W. E. (1993). *The New Economics for Industry, Government, and Education*. MIT Center for Advanced Engineering Study, Cambridge, MA.

Dodson, Bryan (1995). Control charting dependent data: A case study. *Quality Engineering*, **7**, 757-768.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, 14th Edition, 1970. Hafner Press, New York.

Ferrer, Alberto J. and Romero, Rafael (1995). A simple method to study dispersion effects from nonnecessarily replicated data in industrial contexts. *Quality Engineering*, **7**, 747-755.

Hahn, G. J., Hill, W. J., Hoerl, R. W., and Zinkgraf, S. A. (1999). The impact of six sigma improvement—a glimpse into the future of statistics, *The American Statistician*, **53**, 208-215.

Hamada, Michael S., Wilson, Alyson, Reese, C. Shane, and Martz, Harry (2008). *Bayesian Reliability*. Springer, New York.

Holmes, Susan and Ballance, Judy (1994). *Process Improvement Guide*, Second Edition. Air University, Maxwell Air Force Base, AL.

Huzurbazar, Aparna V. (2004). *Flowgraph Models for Multistate Time-to-Event Data.* Wiley, New York.

Ishikawa, Kaoru (1976). *Guide to Quality Control*. Asian Productivity Organization, Tokyo. (Wonderful introduction. Feels like it was written for factory workers.)

Ishikawa, Kaoru (1985). *What is Total Quality Control? The Japanese Way*. Prentice-Hall, Englewood Cliffs, NJ.

Juran, J.M. and Gryna, Frank M. (1993). *Quality Planning and Analysis*, Third Edition. McGraw-Hill, New York.

Kempthorne, O. (1952). *Design and Analysis of Experiments*. Krieger, Huntington, NY.

King, James R. (1995). Binomial statistics and binomial plotting paper: The ugly ducklings of statistics. *Quality Engineering*, **7**, 493-521.

Koopmans, L. H. (1987). *Introduction to Contemporary Statistical Methods*, Second Edition. Duxbury Press, Boston.

Lawson, John S. and Helps, Richard (1996). Detecting undesirable interactions in robust design experiments. *Quality Engineering*, **8**, 465-473.

López-Alvarez, Teresa and Aguirre-Torres, Victor (1997). Improving field performance by sequential experimentation: A successful case study in the chemical industry. *Quality Engineering*, **9**, 391-403.

Lucas, James M. (1994). How to achieve a robust process using response surface methodology. *Journal of Quality Technology*, **26**, 248-260.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, **32**, 1-29.

Luner, Jeffery J. (1994). Achieving continuous improvement with the dual response approach: A demonstration of the Roman catapult. *Quality Engineering*, **6**, 691-705.

McGill, Robert, Tukey, John W., and Larsen, Wayne A. (1978). Variations of Box Plots. *The American Statistician*, **32**, 12-16.

Mahmoud, M.A., Henderson, G.R., Epprecht, E.K., and Woodall, W.H. (2010). Estimating the Standard Deviation in Quality-Control Applications *Journal of Quality Technology*, **42**, 348-357.

Moen, R. D., Nolan, T. W., and Provost, L. P. (1991). *Improving Quality Through Planned Experimentation*. McGraw-Hill, New York.

Montgomery, D.C. and Woodall, W.H. (2008). "An overview of Six Sigma", *International Statistical Revue*, **76**, 329-346.

Myers, Raymond H., Khuri, André I., and Carter, Walter H., Jr., (1989). Response Surface Methodology: 1966-1988. *Technometrics*, **31**, 137-157.

Myers, Raymond H., Khuri, André I., and Vining, Geoffry (1992). Response Surface Alternatives to the Taguchi Robust Parameter Design Approach. *The American Statistician*, **46**, 131-139.

Nair, Vijayan N. (1992). Taguchi's Parameter Design: A Panel Discussion. *Technometrics*, **34**, 127-161.

Nelson, Lloyd S. (1984). The Shewhart control chart – Tests for special causes. *Journal of Qquality Technology*, **16**, 237-239.

Phillips, Aaron R., Jeffries, Rella, Schneider, Jan, and Frankoski, Stanley P. (1998). Using repeatability and reproducibility studies to evaluate a destructive test method. *Quality Engineering*, **10**, 283-290.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305-325.

Pukelsheim, Friedrich (1994). The three sigma rule. *The American Statistician*, **48**, 88-91.

Radson, Darrell and Alwan, Layth C. (1996). Detecting variance reductions using the moving range. *Quality Engineering*, **8**, 165-178.

Ryan, T. P. (1989). *Statistical Methods for Quality Improvement*. John Wiley and Sons, New York.

Sarkar, Ashok (1997). Study on improvement of blow-room performance. *Quality Engineering*, **9**, 529-536.

Schneider, Helmut, Kasperski, William, J., Ledford, Thomas, and Kraushaar, Walter (1996). Control charts for skewed and censored data. *Quality Engineering*, **8**, 263-274.

Schneider, Helmut and Pruett, James M. (1994). Contol charting issues in the process industries. *Quality Engineering*, **6**, 345-373.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. Graduate School of the Department of Agriculture, Washington. Reprint (1986), Dover, New York. (One of my favorite books.)

Shi, Xiquan and Lu, Wei (1994). Bivariate quality control problem under restricted conditions. *Quality Engineering*, **6**, 517-531.

Shina, S. G. (1991). The successful use of the Taguchi method to increase manufacturing process capability. *Quality Engineering*, **3**, 333-349.

Shoemaker, Anne C., Tsui, Kwok-Leung, Wu, C.F. Jeff (1991). Economical Experimentation Methods for Robust Design. *Technometrics*, **33**, 415-427.

Shumway, R. H. (1988). *Applied Statistical Time Series Analysis*. Prentice Hall, Englewood Cliffs, NJ.

Shumway and Stoffer (2011). *Time Series Analysis and Its Applications: With R Examples,* Third Edition. Springer, New York.

Steinberg, D.M., Bisgaard, S., Fisher, N., Hahn, G., Kettering, J., Montgomery, D.C., Doganksoy, N., Gunter, B., Keller-McNulty, S., Meeker, W.Q., Wu, C.F.J. (2008). The Future of Industrial Statistics: A Panel Discussion; *Technometrics*, **50**, 103-127.

Stephens, Matthew P. (1995). Comparison of Robustness of Taguchi's methods with classical ANOVA under conditions of homogeneous variances. *Quality Engineering*, **7**, 147-167.

Stephens, Matthew P. (1996). Effects of heterogeneity of variance on analysis of designed experiments: A comparison of robustness of classical ANOVA with the use of S/N ratios. *Quality Engineering*, **8**, 411-417.

Sullivan, J. H. and Woodall, W. H. (1996). A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, **28** (4), 398-408.

Taguchi, G. (1987). *Systems of Experimental Design*, Vol. 1. Unipub, Kraus International Publications, New York.

Taguchi, Genichi (1987). *Systems of Experimental Design*, Vol. 1 & 2. UNIPUB, White Plains, NY.

Taguchi, Genichi and Wu, Y. (1985). *Introduction to Off-Line Quality Control*. Central Jampan Quality Control Association, Nagaya, Japan.

van Nuland, Y. (1992). ISO 9002 and the circle technique. *Quality Engineering*, **5**, 269-291.

van Nuland, Yves (1994). Do you have doubts about the measurement results, too? *Quality Engineering*, **6** 99-113.

Walton, Mary (1986) *The Deming Management Method*. Perigee, New York. (Delightful introduction to Deming's ideas.)

Woodall, W. H., and Adams, B. M. (1993). The Statistical Design of CUSUM Charts. *Quality Engineering*, **5**(4), 559-570.

Index